STUDIES IN CLASSIFICATION,
DATA ANALYSIS,
AND KNOWLEDGE ORGANIZATION

Classification as a Tool for Research

Hermann Locarek-Junge
Claus Weihs
Editors

Springer

Studies in Classification, Data Analysis, and Knowledge Organization

For further volumes:
http://www.springer.com/series/1564

Hermann Locarek-Junge · Claus Weihs
Editors

Classification as a Tool for Research

Proceedings of the 11th IFCS Biennial Conference and 33rd Annual Conference of the Gesellschaft für Klassifikation e.V., Dresden, March 13–18, 2009

Springer

Editors

Professor Dr. Hermann Locarek-Junge
Chair Finance and Financial Services
Dresden University of Technology
Helmholtzstr. 10
01062 Dresden
Germany
Hermann.Locarek-Junge@tu-dresden.de

Professor Dr. Claus Weihs
Chair Computational Statistics
Dortmund University of Technology
Vogelpothsweg 87
44221 Dortmund
Germany
weihs@statistik.uni-dortmund.de

ISSN 1431-8814
ISBN 978-3-642-10744-3 e-ISBN 978-3-642-10745-0
DOI: 10.1007/978-3-642-10745-0
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2010923661

Cover design: WMX Design, Heidelberg

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume contains revised selected papers from plenary and invited as well as contributed sessions at the 11th Biennial Conference of the International Federation of Classification Societies (IFCS) in combination with the 33rd Annual Conference of the German Classification Society – Gesellschaft für Klassifikation (GfKl), organized by the Faculty of Business Management and Economics at the Technische Universität Dresden in March 2009. The theme of the conference was "Classification as a Tool for Research." The conference encompassed 290 presentations in 100 sessions, including 11 plenary talks and 2 workshops. Moreover, five tutorials took place before the conference. With 357 attendees from 58 countries, the conference provided a very attractive interdisciplinary international forum for discussion and mutual exchange of knowledge.

The chapters in this volume were selected in a second reviewing process after the conference. From the remaining 120 submitted papers, 90 papers were accepted for this volume. In addition to the fundamental methodological areas of Classification and Data Analysis, the volume contains many chapters from a wide range of topics representing typical applications of classification and data analysis methods in Archaeology and Spatial Science, Bio-Sciences, Electronic Data and Web, Finance and Banking, Linguistics, Marketing, Music Science, and Quality Assurance and Engineering.

The editors would like to thank the session organizers for supporting the spread of information about the conference, and for inviting speakers, all reviewers for their timely reports, and Irene Barrios-Kezic and Martina Bihn of Springer-Verlag, Heidelberg, for their support and dedication to the production of this volume.

Moreover, IFCS and GfKl want to thank the Local Organizing Committee, Werner Esswein, Andreas Hilbert, and Hermann Locarek-Junge for this very well-organized conference. We also thank all our supporters – special thanks to Thorsten Klug, Sven Loßagk, Karoline Schönbrunn, Jens Weller, and the student staff at the conference!

Dresden and Dortmund
November 2009

Hermann Locarek-Junge
Claus Weihs

Scientific Program Committee

Chair

Claus Weihs, University of Dortmund, Germany

Members

- David Banks, Duke University, USA
- Vladimir Batagelj, University of Ljubljana, Slovenia
- Patrice Bertrand, Université Paris-Dauphine, France
- Hans-Hermann Bock, RWTH Aachen University, Germany
- Paula Brito, University of Porto, Portugal
- Joachim Buhmann, ETH Zürich, Switzerland
- Andrea Cerioli, University of Parma, Italy
- Eva Ceulemans, Katholieke Universiteit Leuven, Belgium
- Reinhold Decker, Universität Bielefeld, Germany
- Werner Esswein, TU Dresden, Germany
- Bernard Fichet, LIF Marseille, France
- Eugeniusz Gatnar, AE Katowice, Poland
- David Hand, Imperial College London, UK
- Christian Hennig, UCL, UK
- Andreas Hilbert, TU Dresden, Germany
- Tadashi Imaizumi, Tama University Tokyo, Japan
- Krzysztof Jajuga, Wroclaw University of Economics, Poland
- Taerim Lee, Korea National Open University, Korea
- Hermann Locarek-Junge, TU Dresden, Germany
- Buck McMorris, Illinois Institute of Technology, USA
- Fionn Murtagh, University of London, UK
- Akinori Okada, Tama University Tokyo, Japan
- Marieke Timmerman, Rijks Universiteit Groningen, Netherlands
- Maurizio Vichi, Universitá di Roma La Sapienza, Italy

Reviewers (in alphabetical order)

Daniel Baier, Martin Behnisch, Axel Benner, Lynne Billard, Anne-Laure Boulesteix, Alexander Brenning, Hans Burkhardt, Wolfgang Gaul, Patrick J.F. Groenen,Georges Hebrail, Irmela Herzog, Tadashi Imaizumi, Krzysztof Jajuga, Sabine Krolak-Schwerdt, Berthold Lausen, Taerim Lee, Uwe Ligges, Hermann Locarek-Junge, Jorn Mehnen, Andreas Nuernberger, Jozef Pociecha, Axel G. Posluschny, Gunter Ritter, Alex Rogers, Jürgen Rolshoven, Lars Schmidt-Thieme, Wilfried Seidel, Susanne Strahringer, Heike Trautmann, Alfred Ultsch, Rosanna Verde, Claus Weihs

1 Program Sessions and Chairs

1.1 Plenary Sessions

Taylan Cemgil: Hierarchical Bayesian Models for Audio and Music Processing (Chair: Prof. Claus Weihs)
Sanjoy Dasgupta: Performance Guarantees for Hierarchical Clustering (Chair: Prof. Joachim M Buhmann)
Josef Kittler: Data Quality Dependent Decision Making in Pattern Classification (Chair: Prof. Fionn Murtagh)
Lars Schmidt-Thieme: Object Identification (Chair: Prof. Andreas Geyer-Schulz)
Alfred Ultsch: Benchmarking Methods for the Identification of Differentially Expressed Genes (Chair: Dr. Berthold Lausen)
Vincenzo Vinzi: PLS Path Modelling and PLS Regression (Chair: Prof. Yoshio Takane)

1.2 President's Invited Session (Chair: Prof. F.R. McMorris)

Alain Guenoche: String Distances for Complete Genome Phylogeny
Boris Mirkin: Clustering Proteins and Tree-Mapping Evolutionary Events
William Shannon, Elena Deych, Robert Culverhouse: Microarray Dimension Reduction Based on Maximizing Mantel Correlation Coefficients Using a Genetic Algorithm Search Strategy

1.3 Workshops

Sensor Networks (Chairs: Dr. Dimitris Tasoulis, Dr. Niall Adams) (Organizers: Dr. Dimitris K. Tasoulis, Dr. Niall M. Adams, Dr. Alex Rogers)
Bibliothekarischer Workshop (Chairs: Dr. Hans-Joachim Hermes, Dr. Bernd Lorenz) (Organizers: Dr. Hans-Joachim Hermes, Dr. Bernd Lorenz)

1.4 Invited Sessions

Business Informatics (Chairs: Prof. Werner Esswein, Prof. Susanne Strahringer)
Classification Approaches for Symbolic Data (Chair: Prof. Rosanna Verde)
Clustering and Classification (Chair: Prof. Bernard J.E. Fichet)
Clustering in Networks (Chair: Prof. Vladimir Batagelj)

Clustering in Reduced Space
(Chairs: Prof. Eva Ceulemans, Dr. Marieke E. Timmerman)
Correspondence Analysis and Related Methods
(Chair: Prof. Patrick J.F. Groenen)
Data Stream Mining (Chair: Prof. Georges Hebrail)
Graph-Theoretical Methods for Clustering (Chair: Prof. Hans-Hermann Bock)
Information Extraction and Retrieval (Chair: Prof. Lars Schmidt-Thieme)
Modelling Genome Wide Data in Clinical Research I
(Chair: Dr. Berthold Lausen)
Modelling Genome Wide Data in Clinical Research II
(Chair: Prof. Katja Ickstadt)
Model-Based Clustering Methods I (Chair: Dr. Christian Hennig)
Multicriteria Optimization I (Chair: Dr. Heike Trautmann)
Non-Standard Data (Chair: Lynne Billard)
Spatial Classification I (Chair: Prof. Alexander Brenning)
Two-Way Clustering and Applications (Chair: Prof. Hans-Hermann Bock)

1.5 Contributed Sessions

Clustering and Classification

Applied Clustering Methods (Chair: Dr. Andrzej Dudek)
Clustering for Similarity Data (Chair: Prof. Erhard Godehardt)
Clustering: Bias and Stability (Chair: Dr. Christian Hennig)
Comparison/Dynamics in Clustering (Chair: Simona Balbi)
Hierarchical Clustering (Chair: Prof. Maria Paula Brito)
Multiway/Reduced Space Clustering (Chair: Dr. Patrice Bertrand)
Discrimination I (Chair: Prof. Guy Cucumel)
Discrimination II (Chair: Prof. Ulrich Müller-Funk)
Model-Based Clustering Methods II (Chair: Prof. Andrzej Sokolowski)
New Clustering Strategies I (Chair: Dr. Jan W. Owsinski)
New Clustering Strategies II (Chair: Prof. Immanuel M. Bomze)
Selection and Clustering of Variables (Chair: Prof. Jozef Dziechciarz)

Data Analysis Methods

Correspondence Analysis and Related Methods I (Chair: Prof. Jörg Blasius)
Correspondence Analysis and Related Methods II
(Chair: Prof. Michael J. Greenacre)
Correspondence Analysis and Related Methods III
(Chair: Prof. John Gower)
Data Analysis Software (Chair: Prof. Uwe Ligges)

Data Cleaning and Pre-Processing/Ensemble Methods
(Chair: Prof. Eugeniusz Gatnar)
Exploratory Data Analysis I (Chair: Prof. Andrea Cerioli)
Exploratory Data Analysis II (Visualization)
(Chair: Prof. Anthony C. Atkinson)
Exploratory Data Analysis III (Multivariate)
(Chair: Prof. Vincenzo Esposito Vinzi)
Exploratory Data Analysis IV (Chair: Prof. Alfred Ultsch)
Large and Complex Data I (Chair: Prof. Maurizio Vichi)
Large and Complex Data II (Chair: Prof. Tadashi Imaizumi)
Mixture Analysis – Mixture Models in Genetics
(Chair: Prof. Wilfried Seidel)
Mixture Analysis – Mixture Estimation and Model Selection
(Chair: Prof. Angela Montanari)
Non-Gaussian Mixtures (Chair: Prof. Wilfried Seidel)
Non-Standard Data I (Chair: Lynne Billard)
Non-Standard Data II (Chair: Lynne Billard)
Non-Standard Data III (Chair: Lynne Billard)
Pattern Recognition and Machine Learning/Data Analysis
(Chair: Prof. Joachim M. Buhmann)
Regression Mixture Models (Chair: Prof. Angela Montanari)
Visualization of Asymmetry (Chair: Dr. Akinori Okada)
Visualization of Symbolic Data (Chair: Prof. Tadashi Imaizumi)
Visualization I (Chair: Prof. Patrick J.F. Groenen)
Visualization II (Chair: Prof. Patrick J.F. Groenen)

Archaeology and Spatial Science

Archaeology and Historical Geography I
(Chairs: Irmela Herzog, Dr. Tim Kerig)
Archaeology and Historical Geography II
(Chairs: Irmela Herzog, Dr. Tim Kerig)
Spatial Classification II (Chair: Prof. Alexander Brenning)
Spatial Planning (Chair: Dr. Martin Behnisch)

Bio-Sciences

Biostatistics and Bioinformatics - Mult. Tests/Pred. with Genomics Data
(Chair: Prof. Geoffrey J. McLachlan)
Highdimensional Genomics I (Chair: Axel Benner)
Highdimensional Genomics II (Chair: Prof. Iven Van Mechelen)
Medical Health I (Chair: Dr. Berthold Lausen)
Medical Health II (Chair: Prof. Taerim Lee)

Pre-Clinical Development and Biostatistics (Chair: Axel Benner)
SNPs and Genome Analysis (Chair: Prof. Gunter Ritter)

Finance and Banking

Banking and Finance I (Chair: Prof. Ursula Walther)
Banking and Finance II (Chair: Prof. Hermann Locarek-Junge)
Banking and Finance III (Chair: Prof. Matija Mayer-Fiedrich)
Banking and Finance IV (Chair: Prof. Alfred Ultsch)

Linguistics and Text Mining

Linguistics I (Chair: Prof. Jürgen Rolshoven)
Text Mining – Classification (Chair: Prof. Andreas Nuernberger)
Text Mining II (Chair: Prof. Andreas Nuernberger)

Marketing

Marketing and Management Science II (Chair: Prof. Daniel Baier)
Marketing and Management Science III (Chair: Prof. Winfried Steiner)
Marketing and Management Science IV (Chair: Prof. Daniel Baier)
Marketing and Management Science V (Chair: Prof. Winfried Steiner)
Marketing and Management Science VI (Chair: Prof. Reinhold Decker)
Retailing/Direktmarketing (Chair: Prof. Reinhold Decker)

Music Science

Statistical Musicology I (Chair: Prof. Claus Weihs)
Statistical Musicology II (Chair: Prof. Claus Weihs)

Quality Assurance and Engineering

Multicriteria Optimization II (Chair: Dr. Heike Trautmann)
Production Engineering (Chair: Dr. Jorn Mehnen)

Social Sciences

Psychology and Education (Chair: Prof. Sabine Krolak-Schwerdt)
Social Sciences I (Chair: Dr. Akinori Okada)
Social Sciences II (Chair: Prof. Eugeniusz Gatnar)

Web Mining

Web Mining I (Chair: Prof. W. Gaul)
Web Mining II (Chair: Prof. W. Gaul)

Contents

Part III Applications

Archaeology and Spatial Planning

Bio-Sciences

Marketing

Music Science

Contributors

Samah Abu-Assab Chair of Marketing and Innovation Management, Brandenburg University of Technology Cottbus, Postbox 101344, 03013 Cottbus, Germany, samah.assab@tu-cottbus.de

Shotaro Akaho The National Institute of Advanced Industrial Science and Technology, 1-1-1, Umezono, Tsukuba, 305-8568, Japan, s.akaho@aist.go.jp

Ulas Akkucuk Department of Management, Bogazici University, Istanbul, Turkey, ulas.akkucuk@boun.edu.tr

Gerhard Arminger Schumpeter School of Business and Economics, University of Wuppertal, Gaußstr. 20, 42097 Wuppertal, Germany, arminger@statistik.uni-wuppertal.de

Markus Arndt European Patent Office, Erhardt Street 27, 80649 Munich, Germany, marndt@epo.org

Ulrich Arndt data2knowledge GmbH, Wilhelm-Umbach-Street 12, 63225 Langen, Germany, ulrich.arndt@data2knowledge.de

Javier Arroyo Dpto. de Ingeniería del Software e Inteligencia Artificial, Universidad Complutense de Madrid. Prof. José García Santesmases s/n 28040 Madrid, Spain, javier.arroyo@fdi.ucm.es

Abdolhadi Darzian Azizi Institut für Entscheidungstheorie und Unternehmensforschung, Karlsruhe University, Karlsruhe, Germany, ah.darzian.azizi@etu.uni-karlsruhe.de

Dunarel Badescu Département d'informatique, Université du Québec à Montréal, C.P. 8888, Succursale Centre-Ville, Montréal QC, Canada H3C 3P8, badescu.dunarel@uqam.ca

Daniel Baier Chair of Marketing and Innovation Management, Brandenburg University of Technology Cottbus, Postbox 101344, 03013 Cottbus, Germany, daniel.baier@tu-cottbus.de

Antonio Balzanella Facoltá di Economia, Dipartimento di Matematica e Statistica, Universitá degliStudi di Napoli Federico II, Via Cinthia, 80138 Napoli, balzanella2@alice.it

Roland Barriot Université de Toulouse UPS Laboratoire de Microbiologie et Génétique Moléculaires, 31000 Toulouse, France, Roland.Barriot@ibcg.biotoul.fr

and

Université Paul Sabatier, CNRS, LMGM, Bat. IBCG, 118, route de Narbonne, 31062 Toulouse cedex 9, France, roland.barriot@ibcg.biotoul.fr

Hans-Georg Bartel Institute for Chemistry, Humboldt University Berlin, Brook-Taylor-Straße 2, 12489 Berlin, Germany, hg.bartel@yahoo.de

Vladimir Batagelj Faculty of Mathematics and Physics, University of Ljubljana, Jadranska 19, 1000 Ljubljana, Slovenia, vladimir.batagelj@fmf.uni-lj.si

Ingor Baumann Lehrstuhl für Werkstofftechnologie, Fakultät Maschinenbau, TU Dortmund, Leonhard-Euler Str. 2, 44227 Dortmund, Germany, ingor.baumann@udo.edu

Claudia Becker School of Law, Economics, and Business, Martin-Luther-University Halle-Wittenberg, 06099 Halle, Germany, claudia.becker@wiwi.uni-halle.de

Martin Behnisch Institute of Historic Building Research and Conservation, ETH Hoenggerberg, HIT H 21.3, 8093 Zurich, Switzerland, Behnisch@arch.ethz.ch

Martin Berger Fraunhofer Institute for Industrial Mathematics (ITWM), Fraunhofer-Platz 1 67663 Kaiserslautern, Germany, martin.berger@itwm.fraunhofer.de

Wolfgang Bessler Center for Finance and Banking, Justus-Liebig University, Licher Strasse 74, 35394 Giessen, Germany, wolfgang.bessler@wirtschaft.uni-giessen.de

Massimo Bilancia Dipartimento di Scienze Statistiche "Carlo Cecchi", Università degli Studi di Bari, Bari, Italy, mabil@dss.uniba.it

Bernd Bischl Faculty of Statistics, Dortmund University of Technology, 44221 Dortmund, Germany, bernd_bischl@gmx.net

Alix Boc Université du Québec à Montréal, Case postale 8888, succursale Centre-ville, Montréal, QC, Canada H3C 3P8, boc.alix@courrier.uqam.ca

Angela Bohn Wirtschaftsuniversität Wien, 1090 Wien, Austria, Angela.Bohn@gmail.com

Giuseppe Bove Dipartimento di Scienze dell'Educazione, Università degli Studi Roma Tre, Rome, Italy, bove@uniroma3.it

Alexander Brenning Department of Geography and Environmental Management, University of Waterloo, 200 University Ave. W., Waterloo, ON, Canada N2L 3G1, brenning@uwaterloo.ca

Tamara Broderick Statistical Laboratory, University of Cambridge, Cambridge, UK, tb361@statslab.cam.ac.uk

Michael Brusch Institute of Business Administration and Economics, Brandenburg University of Technology Cottbus, Postbox 101344, 03013 Cottbus, Germany, m.brusch@tu-cottbus.de

Michael Bücker Fakultät Statistik, Technische Universität Dortmund, 44221 Dortmund, Germany, buecker@statistik.tu-dortmund.de

Francesco Campobasso Dipartimento di Scienze Statistiche "Carlo Cecchi", Università degli Studi di Bari, Bari, Italy, fracampo@dss.uniba.it

Margarida G.M.S. Cardoso Department of Quantitative Methods, ISCTE Business School, Av. das Forças Armadas 1649-026, Lisboa, Portugal, margarida.cardoso@iscte.pt

Gunnar Carlsson Mathematics Department, Stanford University, Stanford, CA, USA, gunnar@math.stanford.edu

J. Douglas Carroll Rutgers Business School, Newark and New Brunswick, NJ, USA, dcarroll@rci.rutgers.edu

Simona Korenjak Černe Faculty of Economics, University of Ljubljana, Kardeljeva pl. 17, 1000 Ljubljana, Slovenia, simona.cerne@ef.uni-lj.si

Marie Chavent Université de Bordeaux, IMB, CNRS, UMR 5251, France

and

INRIA Bordeaux Sud-Ouest, CQFD team, France, chavent@math.u-bordeaux1.fr

Fabrice Clérot Orange Labs, fabrice.clerot@orange-ft.com

Pedro Contreras Department of Computer Science, Royal Holloway, University of London, 57 Egham Hill, Egham TW20 OEX, England, pedro@cs.rhul.ac.uk

Robert Culverhouse Washington University School of Medicine, St. Louis, MO, USA, rculverh@wustl.edu

Sanjoy Dasgupta University of California, San Diego, CA, USA, dasgupta@cs.ucsd.edu

André C.P.L.F. de Carvalho Department of Computer Science, ICMC, University of São Paulo, Av. Trabalhador Sãocarlense, 400, CEP 13560-970, São Carlos, SP, Brazil, andre@icmc.usp.br

Sabina Denkowska Department of Statistics, Cracow University of Economics, Cracow, Poland, sabina.denkowska@uek.krakow.pl

Mark de Rooij Leiden University Institute for Psychological Research, Leiden, The Netherlands, rooijm@fsw.leidenununiv.nl

Maria Rosaria D'Esposito Department of Economics and Statistics, University of Salerno, Salerno, Italy, mdesposi@unisa.it

Elena Deych Washington University School of Medicine, St. Louis, MO, USA, EDEYCH@dom.wustl.edu

Abdoulaye Baniré Diallo Département d'informatique, Université du Québec à Montréal, C.P. 8888, Succursale Centre-Ville, Montréal QC, Canada H3C 3P8, diallo.abdoulaye@uqam.ca

and

McGill Centre for Bioinformatics and School of Computer Science, McGill University, 3775 University Street, Montréal QC, Canada H3A 2B4

Sonia Pértega Díaz Unidad de Epidemiología Clínica y Bioestadística, Hospital de A Coruña, As Xubias, 84, 15006 A Coruña, Spain, Sonia.Pertega.Diaz@sergas.es

Anna Maria Di Sciullo Université du Québec à Montréal, Case postale 8888, succursale Centre-ville, Montréal, QC, Canada H3C 3P8, di sciullo.anne-marie@uqam.ca

Christian Dittmar Fraunhofer Institute of Digital Media Technology (IDMT), Ehrenbergstr. 31, 98693 Ilmenau, Germany, dmr@idmt.fraunhofer.de

Stephan Dlugosz ZEW Centre for European Economic Research, Mannheim, Germany, dlugosz@zew.de

Jens Dolata Head Office for Cultural Heritage Rhineland-Palatinate (GDKE), Große Langgasse 29, 55116 Mainz, Germany, dolata@ziegelforschung.de

Andrzej Dudek Wrocław University of Economics, Nowowiejska 3, 58-500 Jelenia Góra, Poland, andrzej.dudek@ue.wroc.pl

Ehtibar N. Dzhafarov Department of Psychological Sciences, Purdue University, West Lafayette, IN, USA, ehtibar@purdue.edu

Józef Dziechciarz University of Economics, Wrocław, Poland, jozef.dziechciarz@ue.wroc.pl

Marta Dziechciarz University of Economics, Wrocław, Poland, marta.dziechciarz@ue.wroc.pl

Said Esber Chair of Marketing and Innovation Management, Brandenburg University of Technology Cottbus, Postbox 101344, 03013 Cottbus, Germany, esbersai@tu-cottbus.de

Katti Faceli Federal University of São Carlos, Campus Sorocaba, Rodovia João Leme dos Santos, Km 110, 18052-780, Sorocaba, SP, Brazil, katti@ufscar.br

Annarita Fanizzi Dipartimento di Scienze Statistiche "Carlo Cecchi", Università degli Studi di Bari, Bari, Italy, a.fanizzi@dss.uniba.it

Ingo Feinerer Technische Universität Wien, 1040 Wien, Austria, Feinerer@dbai.tuwien.ac.at

Raphaël Féraud Raphaël Féraud, Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion, France. raphael.feraud@orange-ftgroup.com

Ana Sousa Ferreira LEAD, FPCE, University of Lisbon, Alameda da Universidade, 1649-013 Lisboa, Portugal

and

CEAUL, Multivariate Data Analysis and Modelling Project, Lisboa, Portugal, asferreira@fpce.ul.pt

Gwennaele Fichant Université de Toulouse, UPS, Laboratoire de Microbiologie et Génétique Moléculaires, 31000 Toulouse, France, Gwennaele.Fichant@ibcg.biotoul.fr

and

Université Paul Sabatier, CNRS, LMGM, Bat. IBCG, 118, route de Narbonne, 31062 Toulouse cedex 9, France, gwennaele.fichant@ibcg.biotoul.fr

Karl A. Froeschl University of Vienna, Dr.-Karl-Lueger-Ring 1, 1010 Vienna, Austria, Karl.A.Froeschl@univie.ac.at

Philippe Gambette L.I.R.M.M., UMR CNRS 5506, Université Montpellier 2, Montpellier, France, gambette@lirmm.fr

Dominic Gastes Institute of Decision Theory and Operations Research, University of Karlsruhe, Kaiserstrasse 12, 76131 Karlsruhe, Germany, dominic.gastes@wiwi.uni-karlsruhe.de

Claire Gaugain Université de Toulouse, UPS, Laboratoire de Microbiologie et Génétique Moléculaires, 31000 Toulouse, France, claire_gaugain@yahoo.fr

and

Université Paul Sabatier, CNRS, LMGM, Bat. IBCG, 118, route de Narbonne, 31062 Toulouse cedex 9, France, claire.gaugain@ibcg.biotoul.fr

Wolfgang Gaul Institute of Decision Theory and Operations Research, University of Karlsruhe, Kaiserstrasse 12, 76131 Karlsruhe, Germany, wolfgang.gaul@wiwi.uni-karlsruhe.de

Tim Gollub Faculty of Media, Media Systems, Bauhaus-Universität Weimar, Weimar, Germany, Tim.Gollub@uni-weimar.de

Javier González Universidad Carlos III de Madrid, c/Madrid 126, 28903 Getafe, Spain, javier.gonzalez@uc3m.es

Pascal Gouzien Orange Labs, pascal.gouzien@orange-ftgroup.com

Robert B. Gramacy Statistical Laboratory, University of Cambridge, Cambridge, UK, bobby@statslab.cam.ac.uk

Michael Greenacre Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain, michael@upf.es

Alexander Gribov Department of Computer Oriented Statistics and Data Analysis, Institute of Mathematics, University of Augsburg, Augsburg, Germany, alexander.gribov@math.uni-augsburg.de

Matthias Gruhne Fraunhofer Institute of Digital Media Technology (IDMT), Ilmenau, Germany, matthias.gruhne@idmt.fraunhofer.de

Alain Guénoche IML, CNRS, 163 Avenue de Luminy, Marseille, France, guenoche@iml.univ-mrs.fr

Frédéric Guyon MTI, INSERM-Université Denis Diderot, 36 rue Hélène Brion, Paris, France, frederic.guyon@univ-paris-diderot.fr

Edgar Haimerl Institut für Romanistik, Universität Salzburg, Akademiestraße 24, 5020 Salzburg, Austria, Edgar@Haimerl.eu

Tamas Harczos Fraunhofer Institute of Digital Media Technology (IDMT), Ilmenau, Germany, tamas.harczos@idmt.fraunhofer.de

Christian Hennig Department of Statistical Science, UCL, Gower Street, London WC1E 6BT, UK, chrish@stats.ucl.ac.uk

Julian Holler Center for Finance and Banking, Justus-Liebig University, Licher Strasse 74,35394 Giessen, Germany, julian.holler@wirtschaft.uni-giessen.de

Kurt Hornik Vienna University of Economics and Business, Augasse 2-6, 1090 Vienna, Austria, Kurt.Hornik@wu.ac.at

Milan Hronsky EC3 – E-Commerce Competence Center, Vorlaufstrasse 5/6, 1010 Vienna, Austria, Milan.Hronsky@ec3.at

Katja Ickstadt Faculty of Statistics, TU Dortmund and SFB 475, Dortmund, Germany, ickstadt@statistik.uni-dortmund.de

Tudor Ionescu IKE, Universität Stuttgart, Stuttgart, Germany, tudor.ionescu@ike.uni-stuttgart.de

Antonio Irpino Department of European and Mediterranean Studies, Second University of Naples, Via del Setificio 15, 81100, Caserta, Italy, antonio.irpino@unina2.it

Nataša Kejžar Faculty of Social Sciences, University of Ljubljana, Kardeljeva pl. 5, 1000 Ljubljana, Slovenia, natasa.kejzar@fdv.uni-lj.si

Lidia Khmylko Natural Language Systems Group, University of Hamburg, Vogt-Kölln-Straße 30, 22527 Hamburg, Germany, khmylko@informatik.uni-hamburg.de

Thomas Kiefer Technische Universität Dortmund, Fakultät Statistik, D-44221 Dortmund, Germany, kiefer@statistik.tu-dortmund.de

Josef Kittler Centre for Vision, Speech and Signal Processing University of Surrey, Guildford GU2 7XH, UK, J/Kittler@surrey.ac.uk

Martin Klaus SVI Endowed Chair for International Direct Marketing, DMCC Dialog Marketing Competence Center, University of Kassel, Kassel, Germany, mklaus@wirtschaft.uni-kassel.de

Frank Klefenz Fraunhofer Institute of Digital Media Technology (IDMT), Ilmenau, Germany, frank.klefenz@idmt.fraunhofer.de

Gerd Kopp Lehrstuhl Computergestützte Statistik, Fakultät Statistik, TU Dortmund, Vogelpothsweg 87, 44227 Dortmund, Germany, g.kopp@gmx.net

Anne Krampe Faculty of Statistics, TU Dortmund University, Dortmund, Germany, anne.krampe@uni-dortmund.de

Sebastian Krey Faculty of Statistics, Dortmund University of Technology, 44221 Dortmund, Germany, krey@statistik.tu-dortmund.de

Artus Krohn-Grimberghe Institute of Computer Science, Information Systems and Machine Learning Lab, University of Hildesheim, Germany, artus@ismll.de

Rudolf Kruse Otto-von-Guericke-Universität Magdeburg, Magdeburg, Germany, kruse@iws.cs.uni-magdeburg.de

Vanessa Kuentz Université de Bordeaux, IMB, CNRS, UMR 5251, France

and

INRIA Bordeaux Sud-Ouest, CQFD team, France, kuentz@math.u-bordeaux1.fr

Karl-Heinz Küfer Fraunhofer Institute for Industrial Mathematics (ITWM), Fraunhofer-Platz 1 67663 Kaiserslautern, Germany, karl-heinz.kuefer@itwm.fraunhofer.de

Mirko Kühne Chair of Marketing and Innovation Management, Brandenburg University of Technology Cottbus, Postbox 101344, 03013 Cottbus, Germany, kuehnmi1@tu-cottbus.de

Sonja Kuhnt Faculty of Statistics, TU Dortmund University, Dortmund, Germany, kuhnt@statistik.tu-dortmund.de

J. Kunert Department of Statistics, Technische Universität Dortmund, Dortmund, Germany, kunert@statistik.tu-dortmund.de

Rie Kuwabara Department of Pediatrics of Yokohama Seibu Hospital, Collaboration of Unit of Medical Statistics, Institute of Radioisotope Research, St. Marianna University School of Medicine, Kawasaki 216-8511, Japan

L. Kwiatkowski Department of Mechanical Engineering, Technische Universität Dortmund, Dortmund, Germany, lukas.kwiatkowski@iul.tu-dortmund.de

Uwe Ligges Fakultät Statistik, Technische Universität Dortmund, 44221 Dortmund, Germany, ligges@statistik.tu-dortmund.de

Hermann Locarek-Junge Lehrstuhl für Finanzwirtschaft und Finanzdienstleistungen, Technische Universität Dresden, Dresden, Germany, locarekj@finance.wiwi.tu-dresden.de

Sandra Loureiro University of Aveiro, Campus de Santiago, 3810-193 Aveiro, Portugal, sandra.loureiro@ua.pt

Hanna Lukashevich Fraunhofer IDMT, Ehrenbergstr. 31, 98693 Ilmenau, Germany, dmr@idmt.fraunhofer.de

Patrick Mair Wirtschaftsuniversität Wien, 1090 Wien, Austria, Patrick.Mair@wu.ac.at

Vladimir Makarenkov Laboratoire de bioinformatique, Département d'informatique, Université du Québec à Montréal, C.P. 8888, Succursale Centre-Ville, Montréal QC, Canada H3C 3P8, makarenkov.vladimir@uqam.ca

Waqas Ahmed Malik Department of Computer Oriented Statistics and Data Analysis, Institute of Mathematics, University of Augsburg, Augsburg, Germany, malik@math.uni-augsburg.de

Angelos Markos Department of Applied Informatics, University of Macedonia, Macedonia, Greece, amarkos@uom.gr

Cherif Mballo Laboratoire de bioinformatique, Departement d'informatique, UQAM, C.P. 8888 Succursale Centre-Ville, Montreal, QC, Canada H3C 3P8, mballo.cherif@courrier.uqam.ca

Geoffrey J. McLachlan Department of Mathematics, University of Queensland, Australia

and

Institute for Molecular Bioscience, University of Queensland, Australia, gjm@maths.uq.edu.au

Facundo Mémoli Mathematics Department, Stanford University, Stanford, CA, USA, memoli@math.stanford.edu

George Menexes Lab of Agronomy, School of Agriculture, Aristotle University of Thessaloniki, Thessaloniki, Greece, gmenexes@uom.gr

Wolfgang Menzel Natural Language Systems Group, University of Hamburg, Vogt-Kölln-Straße 30, 22527 Hamburg, Germany, menzel@informatik.uni-hamburg.de

Junichi Mimaya The Research Committee for the National Surveillance on Coagulation Disorders in Japan, Japan, m2taki@marianna-u.ac.jp

Boris Mirkin School of Computer Science, Birkbeck University of London, Malet Street, London, WC1 7HX, UK, mirkin@dcs.bbk.ac.uk

and

Department of Applied Mathematics, Higher School of Economics, Kirpichnaya 33/5, Moscow, Russian Federation, bmirkin@yandex.ru

Magdalena Missler-Behr Chair of Planning and Innovation Management, Brandenburg University of Technology Cottbus, Konrad-Wachsmann-Allee 1, 03046 Cottbus, Germany, magdalena.missler-behr@tu-cottbus.de

Kei Miyazaki Department of Cognitive and Behavioral Science, The University of Tokyo, Komaba 3-8-1, Meguro-ku, Tokyo 153-8902, Japan, miyazaki.behaviormetrics@gmail.com

Hans-Joachim Mucha Weierstrass Institute for Applied Analysis and Stochastics (WIAS), 10117 Berlin, Germany, mucha@wias-berlin.de

Tina Müller Faculty of Statistics, TU Dortmund and SFB 475, Dortmund, Germany, tmueller@statistik.tu-dortmund.de

Alberto Muñoz Universidad Carlos III de Madrid, c/Madrid 126, 28903 Getafe, Spain, alberto.munoz@uc3m.es, tina.mueller@uni-dortmund.de

Fionn Murtagh Department of Computer Science, Royal Holloway, University of London, 57 Egham Hill, Egham TW20 OEX, England

and

Science Foundation Ireland, Wilton Place, Dublin 2, Ireland, fmurtagh@acm.org

Atsuho Nakayama Faculty of Economics, Nagasaki University, 4-2-1 Katafuchi, Nagasaki 850-8506, Japan, atsuho@nagasaki-u.ac.jp

Alexandros Nanopoulos Institute of Computer Science, Information Systems and Machine Learning Lab, University of Hildesheim, Germany, nanopoulos@ismll.de

Boris Naujoks Log!n GmbH, Schwelm, Germany, Boris.Naujoks@login-online.de

Monique Noirhomme-Fraiture University of Namur, Namur, Belgium, monique.noirhomme@fundp.ac.be

Rebecca Nugent Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, USA, rnugent@stat.cmu.edu

Takashi Oga Chiba University, 1-33 Yayoi-Cho, Inage-Ku, Chiba, 263-8522, Japan, ohga@le.chiba-u.ac.jp

Akinori Okada Graduate School of Management and Information Sciences, Tama University, Tokyo, Japan, okada@tama.ac.jp

Masato Okada Nara Institute of Science and Technology, 8916-5, Takayama-cho, Ikoma, Nara, 630-0192, Japan

and

The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, 277-8561, Japan, okada@k.u-tokyo.ac.jp

Shinichiro Omachi Tohoku University, 6-6-05 Aoba, Aramaki, Aoba-ku, Sendai, 980-8579, Japan, machi@ecei.tohoku.ac.jp

Alexia Pacheco Costarican Institute of Electricity, San José, Costa Rica, apacheco@ice.go.cr

Iannis Papadimitriou Department of Applied Informatics, University of Macedonia, Greece, iannis@uom.gr

Marcin Pełka Department of Econometrics and Computer Science, Wroclaw University of Economics, Wroclaw, Poland, marcin.pelka@ue.wroc.pl

Laura Poggio The Macaulay Land use Research Institute, Aberdeen, UK, l.poggio@macaulay.ac.uk

Norman Poh Centre for Vision, Speech and Signal Processing University of Surrey, Guildford GU2 7XH, UK, N.Poh@surrey.ac.uk

Wolfgang Polasek Institute for Advanced Studies, Stumpergasse 56, 1060, Vienna, Austria, polasek@ihs.ac.at

Mike Preuss Chair of Algorithm Engineering, TU Dortmund University, Dortmund, Germany, mike.preuss@tu-dortmund.de

Anita Prinzie Manchester Business School, Marketing Group, Booth Street West, Manchester M15 6PB, UK, anita.prinzie@ugent.be

Klaudia Przybysz University of Economics, Wrocław, Poland, klaudia.przybysz@ue.wroc.pl

Antonio Punzo Dipartimento di Economia e Metodi Quantitativi, Università di Catania, Catania, Italy, antonio.punzo@unict.it

Yves Quentin Université de Toulouse, UPS, Laboratoire de Microbiologie et Génétique Moléculaires, 31000 Toulouse, France, Yves.Quentin@ibcg.biotoul.fr

and

Université Paul Sabatier, CNRS, LMGM, Bat. IBCG, 118, route de Narbonne, 31062 Toulouse cedex 9, France, yves.quentin@ibcg.biotoul.fr

Giancarlo Ragozini Department of Sociology, Federico II University of Naples, Naples, Italy, giragoz@unina.it

Oldemar Rodríguez School of Mathematics, University of Costa Rica, San José, Costa Rica, oldemar.rodriguez@ucr.ac.cr

Elvira Romano Facoltá di Studi Politici, Dipartimento di Studi Europei e Mediterranei, Seconda Universitá degli Studi di Napoli, Via del Setificio 15, 81100 Caserta, Italy, elvira.romano@unina2.it

Dorota Rozmus Department of Statistics, Katowice University of Economics, Bogucicka 14, 40-226 Katowice, Poland, drozmus@ae.katowice.pl

Georg Ruß Otto-von-Guericke-Universität Magdeburg, Magdeburg, Germany, russ@iws.cs.uni-magdeburg.de

Silvina Santana University of Aveiro, Campus de Santiago, 3810193 Aveiro, Portugal, silvina.santana@ua.pt

Jérôme Saracco Université de Bordeaux, IMB, CNRS, UMR 5251, France
and
INRIA Bordeaux Sud-Ouest, CQFD team, France
and
Université Montesquieu Bordeaux IV, GREThA, CNRS, UMR 5113, France, jerome.saracco@u-bordeaux4.fr

Anatol Sargin Fakultät Statistik, Technische Universität Dortmund, D-44221 Dortmund, Germany, sargin@statistik.tu-dortmund.de

Leif Scheuermann Max-Weber-Kolleg, Universität Erfurt, Erfurt, Germany, leif.scheuermann@gmail.com

Julia Schiffner Faculty of Statistics, TU Dortmund and SFB 475, Dortmund, Germany, schiffner@statistik.tu-dortmund.de

P. Schmelzer Department of Mechanical Engineering, Technische Universität Dortmund, Dortmund, Germany, paul.schmelzer@iul.tu-dortmund.de

Sabine Schmidt Chair of Planning and Innovation Management, Brandenburg University of Technology Cottbus, Konrad-Wachsmann-Allee 1, 03046 Cottbus, Germany, schmidts@tu-cottbus.de

Martin Schneider Martin-Luther-Universität Halle-Wittenberg, Halle, Germany, schneider@landw.uni-halle.de

Michael Schröder Fraunhofer Institute for Industrial Mathematics (ITWM), Fraunhofer-Platz 1 67663 Kaiserslautern, Germany, michael.schroeder@iwtm.fraunhofer.de

Alexandra Schwarz Schumpeter School of Business and Economics, University of Wuppertal, Gaußstr. 20, 42097 Wuppertal, Germany, schwarz@statistik.uni-wuppertal.de

Holger Schwender Faculty of Statistics, TU Dortmund and SFB 475, Dortmund, Germany, holger.schwender@tu-dortmund.de

Martin Seim Center for Finance and Banking, Justus-Liebig University, Licher Strasse 74, 35394 Giessen, Germany, martin.seim@wirtschaft.uni-giessen.de

William D. Shannon Washington University School of Medicine, St. Louis, MO, USA, wshannon@wustl.edu

Kazuo Shigemasu Department of Psychology, Teikyo University, Otsuka 359, Hachioji-shi, Tokyo 192, Japan, kshige@bayes.c.u-tokyo.ac.jp

Toshio Shimokawa University of Yamanashi, Yamanashi, Japan, shimokawa@yamanashi.ac.jp

Akira Shirahata The Research Committee for the National Surveillance on Coagulation Disorders in Japan, Japan, m2taki@marianna-u.ac.jp

A. Pedro Duarte Silva Faculdade de Economia e Gestão & CEGE, Universidade Católica Portuguesa at Porto, Rua Diogo Botelho, 1327, 4169-005 Porto, Portugal, psilva@porto.ucp.pt

Pierre Soille Joint Research Centre, European Commission, Ispra, Italy, pierre.soille@jrc.it

Andrzej Sokolowski Department of Statistics, Cracow University of Economics, Cracow, Poland, sokolows@uek.krakow.pl

Benno Stein Faculty of Media, Media Systems, Bauhaus-Universität Weimar, Germany, Benno.Stein@uni-weimar.de

S. Straatmann Department of Statistics, Technische Universität Dortmund, Dortmund, Germany, straatmann@statistik.tu-dortmund.de

Eva Stüber Institute of Business Administration and Economics, Brandenburg University of Technology Cottbus, Postbox 101344, 03013 Cottbus, Germany, eva.stueber@tu-cottbus.de

Werner Stuetzle Department of Statistics, University of Washington, Seattle, WA, USA, wxs@u.washington.edu

Gero Szepannek Faculty of Statistics, Dortmund University of Technology, 44221 Dortmund, Germany, szepannek@statistik.tu-dortmund.de

Masashi Taki Department of Pediatrics of Yokohama Seibu Hospital, Collaboration of Unit of Medical Statistics, Institute of Radioisotope Research, St. Marianna University School of Medicine, Kawasaki 216-8511, Japan, m2taki@marianna-u.ac.jp

Shinobu Tatsunami Department of Pediatrics of Yokohama Seibu Hospital, Collaboration of Unit of Medical Statistics, Institute of Radioisotope Research, St. Marianna University School of Medicine, Kawasaki 216-8511, Japan, s2tatsu@marianna-u.ac.jp

Stefan Theußl Wirtschaftsuniversität Wien, 1090 Wien, Austria, Stefan.Theussl@wu.ac.at

Dirk Thorleuchter Fraunhofer INT, 53879 Euskirchen, Appelsgarten 2, Germany, Dirk.Thorleuchter@int.fraunhofer.de

Wolfgang Tillmann Lehrstuhl für Werkstofftechnologie, Fakultät Maschinenbau, TU Dortmund, Leonhard-Euler Str. 2, 44227 Dortmund, Germany, wolfgang.tillmann@udo.edu

Heike Trautmann Statistics Faculty, TU Dortmund University, Dortmund, Germany, heike.trautmann@statistik.uni-dortmund.de

Mitsuhiro Tsuji Kansai University, Osaka, Japan, tsuji@kansai-u.ac.jp

Takahiko Ueno Department of Pediatrics of Yokohama Seibu Hospital, Collaboration of Unit of Medical Statistics, Institute of Radioisotope Research, St. Marianna University School of Medicine, Kawasaki 216-8511, Japan

Alfred Ultsch Datenbionic Research Group, Hans-Meerwein-Strasse, Philipps-University Marburg, 35032 Marburg, Germany, Ultsch@Mathematik.Uni-Marburg.de, ultsch@informatik.uni-marburg.de

Ali Ünlü Institute of Mathematics, University of Augsburg, 86135 Augsburg, Germany, ali.uenlue@math.uni-augsburg.de

Antony Unwin Department of Computer Oriented Statistics and Data Analysis, Institute of Mathematics, University of Augsburg, Germany, unwin@math.uni-augsburg.de

Dirk Van den Poel Faculty of Economics and Business Administration, Ghent University, 9000 Gent, Tweekerkenstraat 2, Belgium, dirk.vandenpoel@ugent.be

Rosanna Verde Department of European and Mediterranean Studies, Second University of Naples, Via del Setificio 15, 81100, Caserta, Italy, rosanna.verde@unina2.it

Jean Véronis L.I.F., UMR CNRS 6166, Université de Provence, France, jean@veronis.fr

José A. Vilar Departamento de Matematicas, Universidade de A Coruña, Campus de Elviña s/n, 15071 A Coruña, Spain, eijoseba@udc.es

Domenico Vistocco Department of Economics, University of Cassino, Cassino, Italy, vistocco@unicas.it

Evelina Vogli Lehrstuhl für Werkstofftechnologie, Fakultät Maschinenbau, TU Dortmund, Leonhard-Euler Str. 2, 44227 Dortmund, Germany, evelina.vogli@udo.edu

Peter Wagner Martin-Luther-Universität Halle-Wittenberg, Halle, Germany, wagner@landw.uni-halle.de

Ralf Wagner SVI Endowed Chair for International Direct Marketing, DMCC Dialog Marketing Competence Center, University of Kassel, Kassel, Germany, rwagner@wirtschaft.uni-kassel.de

Norbert Walchhofer EC3 E-Commerce Competence Center, 1010 Wien, Austria, Norbert.Walchhofer@ec3.at

Marek Walesiak Wrocław University of Economics, Nowowiejska 3, 58-500 Jelenia Góra, Poland, marek.walesiak@ue.wroc.pl

Kazuho Watanabe Nara Institute of Science and Technology, 8916-5, Takayama-cho, Ikoma, Nara, 630-0192, Japan, wkazuho@is.naist.jp

Claus Weihs Lehrstuhl Computergestützte Statistik, Fakultät Statistik, TU Dortmund, Vogelpothsweg 87, 44227 Dortmund, Germany, weihs@statistik.tu-dortmund.de

S. Wenzel Department of Statistics, Technische Universität Dortmund, Dortmund, Germany, wenzel@statistik.tu-dortmund.de

Leesa Wockner Department of Mathematics, University of Queensland, Brisbane, Australia, l.wockner@uq.edu.au

Satoru Yokoyama Department of Business Administration, Faculty of Economics, Teikyo University, 359 Otsuka Hachiouji City, Tokyo, 192-0395, Japan, satoru@main.teikyo-u.ac.jp

Hsiu-Ting Yu Leiden University Institute for Psychological Research, Leiden, The Netherlands, hyu@fsw.leidenununiv.nl

Part I
(Semi-) Plenary Presentations

Hierarchical Clustering with Performance Guarantees

Sanjoy Dasgupta

Abstract We describe two new algorithms for hierarchical clustering, one that is an alternative to complete linkage, and the other an alternative to the k-d tree. In each case, the new algorithm is shown to admit stronger performance guarantees than the classical scheme it replaces.

1 Introduction

A *hierarchical clustering* is a recursive partitioning of a data set into successively more fine-grained clusterings. At the top of the hierarchy, all points are grouped into a single cluster; and each intermediate level is obtained by splitting the clusters in the level above it.

Hierarchical clustering is a basic primitive of statistics and data analysis. It is used for a variety of purposes, prominent among which are:

1. *Exploratory analysis of data.* Here a typical goal is to discover whether a data set contains meaningful groupings, that is, groupings in which the clusters are clearly defined (usually in the sense of being well separated). Popular algorithms for this kind of analysis are agglomerative bottom-up schemes such as average linkage and complete linkage (Sokal and Sneath 1963).
2. *Tree-based vector quantization* (Gray and Neuhoff 1998). Here the idea is to quantize a large data set, that is, to approximate it with a few representatives such that the quantization error (the typical distance between a data point and its representative) is small. It is irrelevant whether or not the clusters are well-defined. This type of hierarchical clustering arises in audio and video coding, and is often constructed top-down, by repeated application of the k-means algorithm (MacQueen 1967).

S. Dasgupta
University of California, San Diego, CA
e-mail: dasgupta@cs.ucsd.edu

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_1,

3. *Organization of data into a spatial structure.* Here the aim is to facilitate future statistical queries such as nearest-neighbor, or classification, or regression. Such queries generically take time $O(n)$ on a database of n points; but if the points are arranged into a tree, it might be possible to process queries much more efficiently, perhaps even in $O(\log n)$ time. In these applications, the most popular form of hierarchical clustering is probably the k-d tree (Bentley 1975).

These are all important applications, and yet the hierarchical clusterings typically used for them are woefully short on meaningful guarantees. If a data set has well defined clusters, is complete linkage guaranteed to find them? If a set of points can be quantized with very low distortion, will the k-means algorithm necessarily find such a quantization? And are k-d trees really the best trees for speeding up statistical queries? In each case, the answer is no.

This state of affairs is understandable when it is considered that these popular algorithms were developed at a time when data was typically one dimensional. In low dimension, the output of a clustering algorithm can be visually checked to see if it is reasonable, and if it isn't, a different clustering procedure can be used; so it is not urgently necessary to have a mathematical assurance of optimality (or near-optimality) for procedures like complete linkage or k-means. Likewise, a wide range of tree structures are effective for answering statistical queries when data is low dimensional; k-d trees work just fine, and are convenient to implement.

In the present time, data analysis lies at the heart of some of the biggest scientific challenges facing us – such as genomics and climate modeling – but these data are extremely high dimensional. It is no longer possible to visualize them to check whether a clustering is sensible. And many of the procedures that work well in low dimension suffer when applied to high dimensional data, either because the problem of local optima is hugely exacerbated (as in the case of the k-means algorithm) or because they fail to adapt effectively to the geometry of high dimensional space (as in the case of k-d trees). In this new regime, it is crucial to have performance guarantees for clustering.

In this paper, we describe two algorithms for hierarchical clustering that were recently proposed specifically to address the challenges of high-dimensional data analysis. In each case, we start with a performance criterion and find that classical schemes fare badly when subjected to this rigorous test. We then design an alternative with strong performance guarantees. Our first algorithm is a replacement for k-d trees; the second, for complete linkage.

2 A Replacement for k-d Trees

2.1 The Curse of Dimension for Spatial Data Structures

A k-d tree (Bentley 1975) is a spatial data structure that partitions $\mathbb{R}^D$ into hyper-rectangular cells. It is built in a recursive manner, splitting along one coordinate direction at a time (Fig. 1, left). The succession of splits corresponds to a binary tree whose leaves contain the individual cells in $\mathbb{R}^D$.

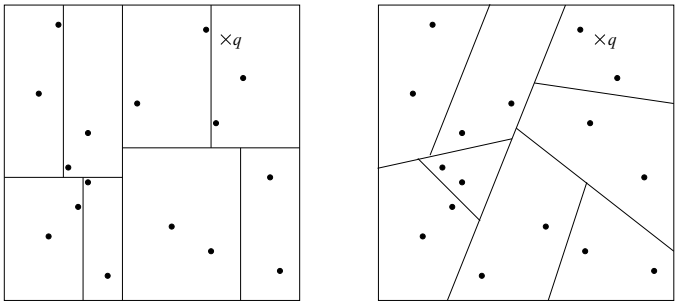


Fig. 1 *Left*: A spatial partitioning of $\mathbb{R}^2$ induced by a k-d tree with three levels. The dots are data points; the cross marks a query point q. *Right*: Partitioning induced by an RP tree

These trees are among the most widely used spatial partitionings in machine learning and statistics. To understand their application, consider Fig. 1(left), and suppose that the dots are points in a database, while the cross is a query point q. The cell containing q, henceforth denoted cell(q), can quickly be identified by moving q down the tree. If the diameter of cell(q) is small (where the diameter is taken to mean the distance between the furthest pair of data points in the cell), then the points in it can be expected to have similar properties, for instance similar labels. In *classification*, q is assigned the majority label in its cell, or the label of its nearest neighbor in the cell. In *regression*, q is assigned the average response value in its cell. In *vector quantization*, q is replaced by the mean of the data points in the cell. Naturally, the statistical theory around k-d trees is centered on *the rate at which the diameter of individual cells drops as you move down the tree*; for details, see page 320 of Devroye et al. (1996).

It is an empirical observation that the usefulness of k-d trees diminishes as the dimension D increases. This can be explained in terms of cell diameter; it is possible to construct a data set in $\mathbb{R}^D$ for which a k-d tree requires D levels in order to halve the cell diameter. In other words, if the data lie in $\mathbb{R}^{1000}$, it could take 1000 levels of the tree to bring the diameter of cells down to half that of the entire data set. This would require $2^{1,000}$ data points!

Here's the construction. Consider $S \subset \mathbb{R}^D$ made up of the coordinate axes between -1 and 1: $S = \bigcup_{i=1}^{D} \{te_i : -1 \leq t \leq 1\}$, where $e_1, \ldots, e_D$ is the canonical basis of $\mathbb{R}^D$. There are many application domains, such as text, in which data is *sparse*; this example is an extreme case. Now, the diameter of S is 2, and it remains 2 even after S is split along one coordinate direction. In fact, it decreases to 1 only after D splits.

Thus k-d trees are susceptible to the same curse of dimensionality that has been the bane of other nonparametric statistical methods.

2.2 Low Dimensional Manifolds and Intrinsic Dimension

A recent positive development in machine learning has been the realization that a lot of data which superficially lie in a very high-dimensional space $\mathbb{R}^D$, actually have low *intrinsic* dimension, in the sense of lying close to a manifold of dimension $d \ll D$. There has been significant interest in algorithms which learn this manifold from data, with the intention that future data can then be transformed into this low-dimensional space, in which standard methods will work well. This field is quite recent and yet the literature on it is already voluminous; early foundational work includes Tenenbaum et al. (2000), Roweis and Saul (2000), and Belkin and Niyogi (2003).

Why is the manifold hypothesis at all reasonable? Suppose, for instance, that you wish to create realistic animations by collecting human motion data and then fitting models to it. A common method for collecting motion data is to have a person wear a skin-tight suit with high contrast reference points printed on it. Video cameras are used to track the 3D trajectories of the reference points as the person is walking or running. In order to ensure good coverage, a typical suit has about $N = 100$ reference points. The position and posture of the body at a particular point of time is represented by a $(3N)$-dimensional vector. However, despite this seeming high dimensionality, the number of degrees of freedom is small, corresponding to the dozen-or-so joint angles in the body. The positions of the reference points are more or less deterministic functions of these joint angles.

To take another example, a speech signal is commonly represented by a high-dimensional time series: the signal is broken into overlapping windows, and a variety of filters are applied within each window. Even richer representations can be obtained by using more filters, or by concatenating vectors corresponding to consecutive windows. Through all this, the intrinsic dimensionality remains small, because the system can be described by a few physical parameters describing the configuration of the speaker's vocal apparatus.

We will adopt a broad notion of intrinsic dimension called the *Assouad* (or *doubling*) *dimension* (Assouad 1983). For any point $x \in \mathbb{R}^D$ and any $r > 0$, let $B(x, r) = \{z : \|x - z\| \leq r\}$ denote the closed ball of radius r centered at x. The Assouad dimension of $S \subset \mathbb{R}^D$ is the smallest integer d such that for any ball $B(x, r) \subset \mathbb{R}^D$, the set $B(x, r) \cap S$ can be covered by 2^d balls of radius $r/2$.

For instance, suppose set S is a line in some high-dimensional space $\mathbb{R}^D$. For any ball B, the intersection $S \cap B$, if nonempty, is a line segment, and it can be covered by exactly two balls of half the radius. Thus the Assouad dimension of S is 1.

A generalization of this argument shows that a d-dimensional affine subspace of $\mathbb{R}^D$ has Assouad dimension $O(d)$. So does a d-dimensional Riemannian submanifold of $\mathbb{R}^D$, subject to a bound on the second fundamental form of the manifold (Dasgupta and Freund 2008). Thus Assouad dimension is more general than the manifold notion we began with.

In fact, it is considerably more general, and also captures *sparsity*, which has recently been a subject of great interest in statistics. For instance, a text document is typically represented as a vector in which each coordinate corresponds to a word and

denotes how often that word occurs within the document. This is an extremely high-dimensional representation if a lot of words are chosen, but it is also sparse – mostly zero – because any given document only contains a tiny subset of the universe of words. It is not hard to show that if S lies in $\mathbb{R}^D$ but has elements with at most d nonzero coordinates, then the Assouad dimension of S is at most $O(d \log D)$.

We are interested in techniques that automatically adapt to intrinsic low dimensional structure without having to explicitly learn this structure. The most obvious first question is, do k-d trees adapt to intrinsic low dimension? The answer is no: the bad example constructed above has an Assouad dimension of just $\log 2D$ (the corresponding set S lies within $B(0, 1)$ and can be covered by $2D$ balls of radius $1/2$.) So we must turn elsewhere.

2.3 Random Projection Trees

Remarkably, a simple variant of k-d trees does adapt to intrinsic dimension. Instead of splitting along coordinate directions at the median, we split along a random direction in S^{D-1} (the unit sphere in $\mathbb{R}^D$), and instead of splitting exactly at the median, we add a small amount of "jitter". We call these *random projection trees* (Fig. 1, right), or RP trees for short. Specifically, for any cell within the tree containing data points (say) S, the splitting rule is determined as follows:

- Choose a random unit direction $v \in \mathbb{R}^D$.
- Pick any $x \in S$; let $y \in S$ be the farthest point from it.
- Choose δ uniformly at random in $[-1, 1] \cdot 6\|x - y\|/\sqrt{D}$.
- All points $\{x \in S : x \cdot v \leq (\text{median}(\{z \cdot v : z \in S\}) + \delta\}$ go to the left subtree; the remainder go to the right.

Suppose an RP tree is built from a data set $S \subset \mathbb{R}^D$, not necessarily finite. If the tree has k levels, then it partitions the space into 2^k cells. We define the *radius* of a cell $C \subset \mathbb{R}^D$ to be the smallest $r > 0$ such that $S \cap C \subset B(x, r)$ for some $x \in C$. Our theorem gives an upper bound on the rate at which the radius of cells in an RP tree decreases as one moves down the tree.

Theorem 1 (Dasgupta and Freund 2008). *There is a constant c_1 with the following property. Suppose an RP tree is built using data set $S \subset \mathbb{R}^D$. Pick any cell C in the RP tree; suppose that $S \cap C$ has Assouad dimension $\leq d$. Then with probability at least $1/2$ (over the randomization in constructing the subtree rooted at C), for every descendant C' which is more than $c_1 d \log d$ levels below C, we have* $\text{radius}(C') \leq \text{radius}(C)/2$.

There is no dependence at all on the extrinsic dimension D.

Since they were introduced, RP trees have been shown to yield algorithms for tree-based vector quantization (Dasgupta and Freund 2009) and regression (Kpotufe 2009) that are adaptive to intrinsic low dimensionality. Also, an efficient scheme for nearest neighbor turns out in retrospect to be using a similar idea (Liu et al. 2004). For experimental work, see Freund et al. (2007).

Open problems

1. An RP tree halves the diameter of cells every $O(d \log d)$ levels; is there an alternative splitting rule that requires just d levels?
2. RP trees and k-d trees are designed for data in Euclidean space. Are there similar constructions (with simple splitting rules) that work in arbitrary metric spaces?
3. What guarantees can be given for query times in nearest neighbor search using RP trees?

3 A Replacement for Complete Linkage

3.1 An Existence Problem for Hierarchical Clustering

We now turn to hierarchical clusterings for exploratory data analysis. Such representations of data have long been a staple of biologists and social scientists, and since the sixties or seventies they have been a standard part of the statistician's toolbox. Their popularity is easy to understand. They require no prior specification of the number of clusters, they permit the data to be understood simultaneously at many levels of granularity, and there are some simple, greedy heuristics that can be used to construct them.

It is very useful to be able to view data at different levels of detail, but the requirement that these clusterings be nested within each other presents some fundamental difficulties. Consider the data set of Fig. 2, consisting of six evenly spaced collinear points in the Euclidean plane. The most commonly used clustering cost functions, such as that of k-means, strive to produce clusters of small radius or diameter. Under such criteria, the best 2-clustering (grouping into two clusters) of this data is unambiguous, as is the best 3-clustering. However, they are hierarchically incompatible. This raises a troubling question: by requiring a hierarchical structure, do we doom ourselves to intermediate clusterings of poor quality?

To rephrase this more constructively, must there always exist a hierarchical clustering in which, for *every* k, the induced k-clustering (grouping into k clusters) is close to the optimal k-clustering under some reasonable cost function? As we have already seen, it is quite possible that the optimal cost-based k-clustering cannot be obtained by merging clusters of the optimal $(k+1)$-clustering. Can they be so far removed that they cannot be reconciled even approximately into a hierarchical structure? We resolve this fundamental existence question via the following result.

Theorem 2 (Dasgupta and Long 2005). *Take the cost of a clustering to be the largest radius of its clusters. Then, any data set in any metric space has a hierarchical clustering in which, for each k, the induced k-clustering has cost at most eight times that of the optimal k-clustering.*

• • • • • •

Fig. 2 What is the best hierarchical clustering for this data set?

Moreover, we have an algorithm for constructing such a hierarchy which is similar in simplicity and efficiency to the popular complete linkage agglomerative clustering algorithm. Complete linkage has the same underlying cost function, but does not admit a similar guarantee.

Theorem 3 (Dasgupta 2009). *For any k, there is a data set for which complete linkage induces k-clusterings whose cost is k times that of the optimal k-clustering.*

3.2 Approximation Algorithms for Clustering

There has been a lot of recent work on the *k-center* and *k-median* problems. In each of these, the input consists of points in a metric space as well as a preordained number of clusters k, and the goal is to find a partition of the points into clusters $C_1, \ldots, C_k$, and also cluster centers $\mu_1, \ldots, \mu_k$ drawn from the metric space, so as to minimize some cost function which is related to the radius of the clusters.

1. *k-center:* Maximum distance from a point to its closest center
2. *k-median:* Average distance from a point to its closest center

Both problems are NP-hard but have simple constant-factor approximation algorithms. For k-center, a two-approximation was found by González (1985), and this is the best approximation factor possible (Feder and Greene 1988). For k-median there have been a series of results; for instance (Arya et al. 2001), achieves an approximation ratio of $6 + \epsilon$, in time $n^{O(1/\epsilon)}$.

What does a constant-factor approximation mean for a clustering problem? Consider the scenario of Fig. 3, set in the Euclidean plane. The solid lines show the real clusters, and the three dots represent the centers of a bad 3-clustering whose cost (in either measure) exceeds that of the true solution by a factor of at least 10. This clustering would therefore not be returned by the approximation algorithms we mentioned. However, EM and k-means regularly fall into local optima of this kind, and practitioners have to take great pains to try to avoid them. In this sense, constant-factor approximations avoid the worst: they are guaranteed to never do too badly. At the same time, the solutions they return can often use some fine-tuning, and local improvement procedures like EM might work well for this.

Although most work on approximation algorithms has focused on flat k-clustering, there is some other work on hierarchies. A different algorithm for the same cost function as ours is given in Charikar et al. (2004); while Plaxton (2003) works with the k-median cost function. More recently, Lin et al. (2006) gives a unifying framework that is able to adapt algorithms for flat clustering to make them hierarchical.

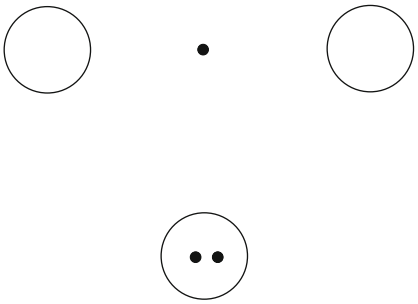


Fig. 3 The circles represent an optimal 3-clustering; all the data points lie within them. The dots are centers of a really bad clustering

Input: n data points with a distance metric $d(\cdot, \cdot)$.
Pick a point and label it 1.
For $i = 2, 3, \ldots, n$
    Find the point furthest from $\{1, 2, \ldots, i-1\}$ and label it i.
    Let $\pi(i) = \arg\min_{j<i} d(i, j)$.
    Let $R_i = d(i, \pi(i))$.

Fig. 4 *Farthest-first traversal* of a data set. Take the distance from a point x to a set S to be $d(x, S) = \min_{y \in S} d(x, y)$

3.3 Farthest-First Traversal

Our algorithm for hierarchical clustering is based upon the *farthest-first traversal* of a set of points, devised by González (1985) as an approximation algorithm for the closely related k-center problem. His use of this traversal for clustering is ingenious, and in fact just a cursory examination of its properties is necessary for his results. For hierarchical clustering, we examine it in greater detail and need to built upon it. Specifically, the farthest-first traversal of n data points yields a sequence of "centers" $\mu_1, \ldots, \mu_n$ such that for any k, the first k of these centers define a k-clustering which is within a factor two of optimal. However, the n clusterings created in this way are not hierarchical. Our main contribution is to demonstrate a simple and elegant way of using the information found by the traversal to create a hierarchical clustering.

The *farthest-first traversal* of a data set starts by picking any data point, then the point furthest from it, then the point furthest from the first two, and so on until k points are obtained. These points are taken as cluster centers and each remaining point is assigned to the closest center. If the distance function is a metric, the resulting clustering is within a factor two of optimal.

Starting with n points in a metric space, number *all* the points in it using a farthest-first traversal (Fig. 4). For any point i, describe its closest neighbor among $1, 2, \ldots, i-1$ as its *parent*, $\pi(i)$. Let R_i be its distance to this parent,

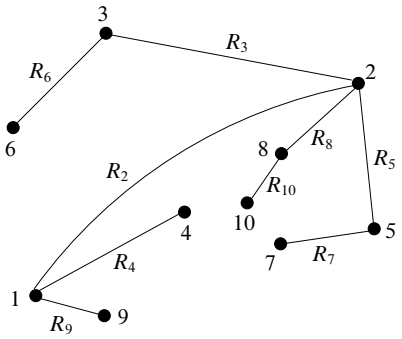


Fig. 5 A farthest-first traversal of ten data points in the plane, under Euclidean distance. The numbering is completely determined by the choice of point number one (and by the method of breaking any ties that arise)

$$R_i \;=\; d(i, \pi(i)) \;=\; d(i, \{1, 2, \ldots, i-1\}).$$

Then $R_1 \geq R_2 \geq R_3 \geq \cdots \geq R_n$. Figure 5 shows an example with a toy data set of ten points.

The algorithm of González uses points $1, 2, \ldots, k$ as centers for a k-clustering. Let $\mathbb{C}_k$ be this clustering; notice that its cost is exactly R_{k+1}.

Theorem 4 (González 1985). *For any k, any k-clustering must have at least one cluster of diameter $\geq R_{k+1}$. Thus, cost$(\mathbb{C}_k) = R_{k+1} \leq 2 \cdot$ cost(optimal k-clustering).*

3.4 A Hierarchical Clustering Algorithm

A farthest-first traversal orders the points so that for any k, the first k points constitute the centers of a near-optimal k-clustering $\mathbb{C}_k$. Unfortunately, the n clusterings defined in this manner are not hierarchical. In Fig. 5 for instance, the 2-clustering clearly puts point 6 in the cluster centered at 1, and point 3 in the cluster centered at 2. However, in the 3-clustering points 3 and 6 are grouped together.

We need a simple scheme for producing a hierarchical clustering starting with a numbering of the data points and an associated parent function π. The tree of Fig. 5 is suggestive. Initially it consists of one connected component: one big cluster. Deleting an edge from the tree breaks this into two connected components, two clusters. Removing another edge will subdivide one of these two clusters, and so on.

Definition A hierarchical clustering $\{\mathbb{C}_1^\rho, \ldots, \mathbb{C}_n^\rho\}$ based on a mapping ρ:

- Pick any function $\rho : \{2, \ldots, n\} \to \{1, 2, \ldots, n\}$ for which $\rho(i) < i$. This property is certainly satisfied by parent function π.

- The graph on nodes $\{1, 2, \ldots, n\}$, with edges $\{(i, \rho(i)) : 2 \leq i \leq n\}$, is a tree. Call it T^ρ.
- For any k, the k-clustering $\mathbb{C}_k^\rho$ is defined as follows.
  - Remove the $k-1$ edges $(2, \rho(2)), \ldots, (k, \rho(k))$ from T^ρ.
  - This leaves k connected components.
  - Each cluster in $\mathbb{C}_k^\rho$ consists of the points in one of these components.

Witness that the clusterings $\{\mathbb{C}_1^\pi, \mathbb{C}_2^\pi, \ldots, \mathbb{C}_n^\pi\}$ *are* hierarchical.

However, the hierarchical clustering generated by T^π might be very poor. To get a sense of what's lacking, look again at Fig. 5. Pick any node k in this tree, remove the edge $(k, \pi(k))$, and consider the connected component containing k. The nodes in this component are grouped together in the k-clustering. The immediate neighbors of k are very close to it – at most R_{k+1} away, and this in turn is at most twice the cost of the optimal k-clustering (recall Theorem 4). But other nodes in this cluster could potentially be much further away.

We will therefore construct an alternative parent function π' whose tree $T^{\pi'}$ has the following property: as you move along any path with increasing node numbers, the edge lengths are bounded by a geometrically decreasing sequence. This immediately rules out the bad effect mentioned above, and as a consequence $\text{cost}(\mathbb{C}_k^{\pi'}) \leq O(1) \cdot \text{cost}(\mathbb{C}_k)$.

We will build π' by viewing the data at certain specific levels of granularity. Let $R = R_2$; this is some rough measure of the span of the data. If we do not care about distances smaller than R, the entire data set can be summarized by the single point $\{1\}$. This is our coarsest view, and we will call it L_0, granularity level zero. Suppose we want a little more detail, but we still don't care about distances less than $R/2$. Then the data can be summarized by L_0 augmented with $L_1 = \{i : R/2 < R_i \leq R\}$. Continuing in this manner, we construct levels $L_0, L_1, L_2, \ldots$ such that every data point is within distance $R/2^j$ of $L_0 \cup L_1 \cup \cdots \cup L_j$.

Earlier we set the parent of i to be its closest neighbor amongst $\{1, 2, \ldots, i-1\}$. We now choose parents from a more restricted set: the closest point *at a lower level of granularity*. The resulting hierarchical clustering algorithm is shown in Fig. 6, and its effect on our earlier example can be seen in Fig. 7. In Dasgupta and Long (2005), it is shown that the algorithm obeys the performance guarantee of Theorem 2.

Open problems

1. It can be shown that any hierarchical scheme for the maximum-radius cost function must have an approximation factor of at least 2. But our algorithm has a factor of 8; can this gap be closed?
2. For complete linkage, it is known that the approximation factor is at least k; is there a matching upper bound?

Input: n data points with a distance metric $d(\cdot, \cdot)$.
Numbering the points
Number the points by farthest-first traversal (Figure 4).
For $i = 2, 3, \ldots, n$, let $R_i = d(i, \{1, 2, \ldots, i-1\})$.
Let $R = R_2$.
Levels of granularity
Lowest level: $L_0 = \{1\}$.
For $j > 1$, $L_j = \{i : R/2^j < R_i \leq R/2^{j-1}\}$.
Hierarchical clustering
Parent function: $\pi'(i) =$ closest point to i at lower level of granularity.
Return the hierarchical clustering corresponding to tree $T^{\pi'}$.

Fig. 6 A hierarchical clustering procedure

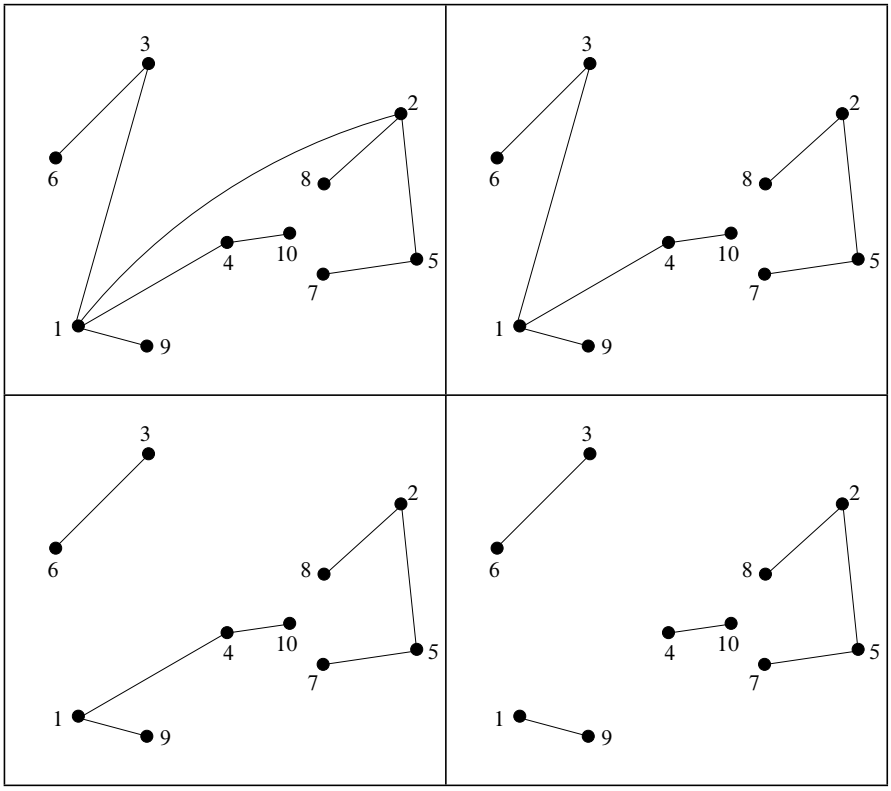


Fig. 7 A continuation of the example of Figure 5. Shown are the 1-, 2-, 3-, and 4-clusterings obtained from the modified parent function π'

References

Arya, V., Garg, N., Khandekar, V., Pandit, V., Meyerson, A., & Munagala, K. (2001). Local search heuristics for k-median and facility location problems. *Proceedings of the 33rd ACM Symposium on the Theory of Computing*, 21–29.

Assouad, P. (1983). Plongements lipschitziens dans r^n. *Bulletin Société Mathématique de France, 111*(4), 429–448.

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation, 15*(6), 1373–1396.

Bentley, J. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM, 18*(9), 509–517.

Charikar, M., Chekuri, C., Feder, T., & Motwani, R. (2004). Incremental clustering and dynamic information retrieval. *SIAM Journal on Computing, 33*(6), 1417–1440.

Dasgupta, S. (2009). What approximation ratio is achieved by complete linkage? *Manuscript*.

Dasgupta, S., & Freund, Y. (2008). Random projection trees and low-dimensional manifolds. *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*. (pp. 537–546). Victoria, British Columbia, Canada.

Dasgupta, S., & Freund, Y. (2009). Random projection trees for vector quantization. *IEEE Transactions on Information Theory, 55*(7).

Dasgupta, S., & Long, P. M. (2005). Performance guarantees for hierarchical clustering. *Journal of Computer and Systems Sciences, 70*(4), 555–569.

Devroye, L., Gyorfi, L., & Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Berlin, Heidelberg, New York: Springer.

Feder, T., & Greene, D. (1988). Optimal algorithms for approximate clustering. *Proceedings of the 20th ACM Symposium on the Theory of Computing* (pp. 434–444). Chicago, Illinois, US.

Freund, Y., Dasgupta, S., Kabra, M., & Verma, N. (2007). Learning the structure of manifolds using random projections. Advances in Neural Information *Processing Systems, 20*, Vancouver, British Columbia, Canada.

González, T. F. (1985). Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science, 38*, 293–306.

Gray, R., & Neuhoff, D. (1998). Quantization. *IEEE Transactions on Information Theory, 44*(6), 2325–2383.

Kpotufe, S. (2009). Escaping the curse of dimensionality with a tree-based regressor. Conference on Learning Theory. Montreal, Quebec, Canada.

Liu, T., Moore, A., Gray, A., & Yang, K. (2004). An investigation of practical approximate nearest neighbor algorithms. *Advances in Neural Information Processing Systems, 17*, (pp. 825–832). Vancouver, British Columbia, Canada.

Lin, G., Nagarajan, C., Rajaraman, R., & Williamson, D. (2006). A general framework for incremental approximation and hierarchical clustering. *17th ACM/SIAM Symposium on Discrete Algorithms*, 1147–1156.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, 1*, 281–297.

Plaxton, G. (2003). Approximation algorithms for hierarchical location problems. *Proceedings of the 35th ACM Symposium on the Theory of Computing*, 40–49.

Roweis, S., & Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science, 290*, 2323–2326.

Sokal, R., & Sneath, P. (1963). *Principles of taxonomy.* New York (San Francisco, CA): Freeman.

Tenenbaum, J., de Silva, V., & Langford, J. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science, 290*(5500), 2319–2323.

Alignment Free String Distances for Phylogeny

Frédéric Guyon and Alain Guénoche

Abstract In this paper, we compare the accuracy of four string distances to recover correct phylogenies of complete genomes . These distances are based on common words shared by raw genomic sequences and do not require preliminary processing steps such as gene identification or sequence alignment. Moreover, they are computable in linear time.

The first distance is based on Maximum Significant Matches. The second is computed from the frequencies of all the words of length k. The third one is based on the Average length of maximum Common Substrings at any position. The last one is based on the Ziv-Lempel compression algorithm.

We describe a simulation process of evolution to generate a set of sequences having evolved according to a random tree topology T. This process allows both base substitutions and fragment insertion/deletion, including horizontal gene transfers. The distances between the generated sequences are computed using the four string formulas and the corresponding trees T' are reconstructed using Neighbor-Joining. Trees T and T' are compared using three topological criteria. These comparisons show that the MSM distance outperforms the others whatever the parameters used to generate sequences.

1 Introduction

More than 800 complete sequences of bacterial genomes are now available at the NCBI and this number is rapidly increasing. Consequently, many recent works deal with phylogenies based on whole genome information rather than on a single or a small number of genes. Whole genome distance computations can be categorized in: (a) frequencies of common words or motifs, (b) presence or absence of sheared homologous genes, (c) gene order along the chromosomes, (d) assembly of several gene trees (see Snel 2005 for more details). The three last categories of methods,

A. Guénoche (✉)
IML, CNRS, 163 Avenue de Luminy, Marseille, France
e-mail: guenoche@iml.univ-mrs.fr

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_2,

imply the identification of orthologous genes that are derived from an ancestral one following a speciation event. This step is often misleading, even for closely related genomes, because genes are subject to duplications, losses and horizontal transfers (*HGT*).

By contrast, category (a) contains distances between genome sequences without gene identification or alignment. Theses distances are based either from the frequencies of DNA words having a fixed length k or from maximal common words (substrings). The usual criticism about them is that the corresponding distances could not be considered as evolutive since they do not derive from a model of sequence evolution. Consequently, the inferred trees are suspicious for phylogeny. We try to assess this statement selecting four alignment free distances computable in linear time: the Maximum Significant Matches (MSM) distance, which improves the Maximum Unique Matches (MUM) distance described in Guyon and Guénoche (2008), a k-word (KW) distance (Qi et al. 2004), the Average Common Substring (ACS) distance (Ulitsky et al. 2006) and one of the compression distances (ZL) defined by Otu and Sayood (2003).

The aim of this paper is to compare these distances according to their accuracy to recover the correct phylogenetic tree, using simulated data. Our paper is organized as follows:

- In Sect. 2, we recall the definitions of the four distances, establishing the minimum length of a common word to be significant.
- In Sect. 3, we describe an evolutionary model including nucleotide substitutions, insertions and deletions of fragments, making large variations in base composition and length. Then, using several topological criteria, we compare the recovered NJ trees to those used to generate sequences.

2 Four Alignment Free Distances

2.1 The MSM Distance

We define a Maximum Significant Match (MSM) as a word that is present in two DNA sequences, which cannot be extended without mismatch and which is not expected to occur by chance. So, we first estimate the minimum length for which a maximal match is *significant*, according to the length and base composition of the two compared sequences.

Let G_1 and G_2 be two DNA sequences of L_1 and L_2 characters over the alphabet $\mathcal{A} = \{\text{A,C,G,T}\}$ and $N_i(\alpha)$ be the number of characters α in genome G_i. We assume that the sequences satisfy an i.i.d model having successive characters sampled independently with distribution $\mu_i(\alpha) = \frac{N_i(\alpha)}{L_i}$ in sequence G_i. Hence, the probability of a character match between two sequences is given by

$$p_{match} = \sum_{\alpha \in \mathcal{A}} \mu_1(\alpha)\mu_2(\alpha).$$

Let N_l be the expected number of common words with a length greater than l; it is given by the limit of the geometric series

$$N_l = (1 - p_{match})^2 L_1 L_2 \sum_{k=l}^{\infty} p_{match}^k = (1 - p_{match}) L_1 L_2 p_{match}^l \qquad (1)$$

We define the significant length denoted l_{min} to be the smallest length such that the expected number of common words larger than l_{min} is lower than 1 in random sequences. From (1)

$$l_{min} \geq -\frac{\log(L_1 L_2 (1 - p_{match}))}{\log(p_{match})}.$$

In practice, to get an integer value L_{sign}, which is sufficient to assert that a common word of such length is unlikely to occur in random sequences, we round up l_{min} to

$$L_{sign} = 1 + \lfloor l_{min} + .5 \rfloor.$$

This value has been tested by simulations and provides better results than $\lceil l_{min} \rceil$. According to this *Lsign* definition, the average MSM number between two random sequences is observed to be lower than .5, whatever are the base compositions and the sequence lengths.

So, a Maximal Significant Matches (MSM) is a maximal common word not smaller than L_{sign}. To define the MSM distance function, we consider the sum of length of these words :

$$D_{MSM}(G_1, G_2) = -\log \frac{\sum |MSM(G_1, G_2)|}{\min\{L_1, L_2\}}$$

When there is no MSM the numerator is set to 1 to avoid infinite distance value.

The MSM identification is performed by a suffix tree which is a very efficient structure for finding all the matches common to two strings. It can be constructed in linear time, using a linear space. For computation, we use the MUMmer suffix tree package developed by Kurtz et al. (2004).

2.2 The k-word distance

Taking into account the frequencies of DNA words to compare genomes is not new (Karlin 1995). The basic idea is to use the frequency vector of all the words of fixed length k present in a sequence. This vector is very easy to compute in linear time, moving a k-width window along the sequence. Usual formulas, such as euclidean or manhattan distances between these vectors, are not very accurate for

precise phylogenetic reconstruction, even when frequencies are corrected to take base composition heterogeneity into account.

In an article, devoted to phylogenetic reconstruction from distances between complete genome sequences, Qi *et al.* have tested a more accurate string distance. The frequencies of all the words of length k are computed but also those of length $k-1$ and $k-2$. Let $F_i(a_1,..a_k)$ be the observed frequency of word $(a_1,..a_k)$ within the G_i sequence, both strands being considered. The expected value, according to a Markov model of order $k-1$, is

$$E_i(a_1,..a_k) = \frac{F_i(a_1,..a_{k-1})F_i(a_2,..a_k)}{F_i(a_2,..a_{k-1})}.$$

Thus the authors do not work anymore with raw frequencies, but with their variations over what is expected. They associate to each genome G_i a vector v_i indexed over all the words of length k, each component being equal to:

$$v_i(a_1,..a_k) = \frac{F_i(a_1,..a_k) - E_i(a_1,..a_k)}{E_i(a_1,..a_k)}.$$

These vectors are compared measuring the cosine value of their angle. A simple normalisation permits to get a distance value in [0,1].

$$KW(G_1,G_2) = (1 - \frac{v_1^\top v_2}{\|v_1\|_2\|v_2\|_2})/2$$

2.3 The ACS distance

The third distance is also based on longest common words between two sequences. It has been introduced by Ulitsky et al. (2006) as the Average length of longest Common Substrings starting at any position in both sequences.

In each position in G_1, a longest word common to G_2 is searched. Let w_i be this word starting in position i in G_1 that can be anywhere in G_2 and let $|w_i|$ be its length. The larger is $\sum_{i=1,..L_1}|w_i|$ the closer is G_1 to G_2. Considering that this sum is increased when L_2 is high, the similarity between G_1 and G_2 is normalised:

$$S(G_1,G_2) = \frac{\sum_{i=1}^{L_1}|w_i|}{L_1\log(L_2)}.$$

As generally $S(G_1,G_2) \neq S(G_2,G_1)$, the ACS distance is defined as the average of the inverse of the two similarity values.

$$ACS(G_1,G_2) = \frac{1}{2}\left(\frac{1}{S(G_1,G_2)} + \frac{1}{S(G_2,G_1)}\right)$$

In the original publication, there is a correction term to insure $ACS(G, G) = 0$, which is not considered here because it tends very quickly to 0. The formula is justified in case the strings were generated by unknown Markov processes. It can be computed in linear time with a suffix tree structure, but the implementation of a suffix array (lexicographical order on suffixes) gives an acceptable time complexity in $O(L \log(L))$ to evaluate a single similarity value.

As it is described, this distance considers only one strand, because it has been applied by the authors to protein sequences. For DNA genomes, we compare G_1 to the both strands of G_2 and so w_i can be on one or the other.

2.4 A Compression Distance

Compression distances are derived from the Kolmogorov complexity theory, considering the smallest size of an automata (program) permitting to generate a sequence. The most regular is the sequence, the shortest is the program. But no procedure can guarantee that an automata has the minimum size. So, most of the researchers use the file compression algorithm due to Ziv and Lempel (1977), which is still intensively used. Its principle is to look for new words in a sequence. It seeks for the longest repeated word starting at the current position and, adding one character, it provides a shortest new word and set the next current position hereafter. This procedure consists in slicing sequence G into consecutive words $G = (g_1|g_2|..|g_p)$ such that $g_i = (a_1..a_k)$ is the shortest word which is not present in prefix $G_{i-1} = (g_1|..|g_{i-1})$ extended with the $k - 1$ characters of g_i. This implies that $(a_1..a_{k-1})$ is present in G before the a_{k-1} position.

Doing so, word g_1 necessarily has just one character a_1, and also is g_2 except if g_2 begins with character a_1, etc. For instance, $G = (acacagtagtcag)$ will be sliced into six words, $(a|c|acag|t|agtc|ag)$, the third being $g_3 = (acag)$ since aca is a previous prefix (in position 1), but $acag$ is not.

The important quantity in the Ziv-Lempel algorithm is the number of words in this decomposition. This function is classically denoted by h. In fact $h(G)$ is the number of shortest new words in G. Here $h(acacagtagtcag) = 5$, since the last word, ag is not new. The h function is intensively used to define the five distances proposed by Otu and Sayood (2003); we retain the last one:

Considering two genomes G_1 and G_2 let $G_1 + G_2$ be the concatenated sequence of them two. It is clear that $h(G_1 + G_2) \leq h(G_1) + h(G_2)$, since the new words found in G_2 after the G_1 slicing can have been previously found.

$$ZL(G_1, G_2) = \frac{h(G_1 + G_2) - h(G_1) + h(G_2 + G_1) - h(G_2)}{h(G_1) + h(G_2)}$$

which corresponds, as the authors say, to the G_2 compression knowing G_1 plus the G_1 compression knowing G_2 devided by the compressions of G_1 and G_2.

These distance values, between 0 and 1, can also be efficiently computed using a suffix-tree as for the MSM distance.

3 Simulations

Sequences are generated according to a tree T, and random mutational events occurring along the edges. The tree shape is selected at random, as the edge's lengths.

3.1 A Simple Evolutionary Model

It depends on four parameters. The first one *Ind* represents the average number of insertions and deletions of DNA fragments in the tree. These indel fragments can occur in any edge and produce sequence length variations. Both losses and gains are equally probable:

- Deletion of a DNA segment at any position, covering at most 1/10-th of the sequence length;
- Insertion of a DNA segment no larger than 1/4-th of the sequence length, at any position. With the same probability, it can be a duplication of the adjacent fragment, as in a tandem repeat, or a fragment taken from another sequence in the tree, simulating an horizontal transfer.

A second parameter Rev allows to fix the average number of reversed fragments between the ancestral sequence and any terminal one. As for the indels, these reversals can arise along any internal edge.

A third parameter, Sub, refers to the percentage of positions in each sequence where a substitution occurred all along the evolutionary process. The number of substitutions between two successive nodes is proportional to the length of this edge and the mutated positions are selected at random. In 3/4 cases it is a transition ($A \leftrightarrow G$ or $T \leftrightarrow C$) and in 1/4 case it is a transversion, such as in a Kimura$_{2p}$ model (1980).

A fourth parameter indicates if the base composition (BC) remains constant or not along the evolutionary process. When $BC = 0$ the substitution rate is the same all over the tree and produces sequences having nearly the same proportion of nucleotides as the ancestral sequence. With $BC = 1$, at each bifurcation, some mutations to A or T on one side, and to G or C on the other side, are inhibited. Consequently, terminal sequences at the end of the subdivisions can present heterogeneous base composition.

This four parameters evolutionary model is used to generate sets of sequences having evolved according to a random phylogenetic tree. It allows generating

sequences having a length varying from one to the double and with a G + C content ranging from 25% to 75%, as it is often observed in real genomes.

3.2 The Simulation Process

To generate random phylogenetic trees, we use the Yule-Harding procedure (1971). Edge lengths are uniformly selected in the range [1, 10], providing large variations. The simulation process consists in:

1. Starting from an ancestral random DNA sequence, the four bases being equiprobable;
2. Generating a set of n terminal sequences (after $n - 2$ internal ones), following a random topology (T) and the evolutionary model described above;
3. Estimating the distance between pairs of terminal sequences, using each of the four distances;
4. Reconstructing the corresponding phylogenetic tree (T') using the neighbor-joining method (Saitou 1987);
5. Comparing T' to T, using three classical criteria:

   - The number, RF of internal edges in T' which are not in T; as both trees have $(2n - 3)$ edges, it is half the Robinson-Foulds distance (Robinson 1981) between X-tree topologies.
   - The number of quadruples *NbQ* which do not have the same topology in both trees; this quantity, divided by the total number of quadruples, is another distance between X-trees, more progressive than the first one (Estabrook 1985).
   - The maximum number of leaves for which the initial and the computed trees are topologically identical. This parameter is classically denoted as the *MAST* value, for Maximum Agreement Sub-Tree (Amir and Keselman 1997). It is a similarity index, bounded by n so, to keep a distance index we edit ($n-MAST$) value, corresponding to the number of taxa to erase to get identical subtree topologies.

   These criteria are independent of the edge lengths. They only compare unrooted phylogenetic trees as they are provided by NJ; for these comparisons T is considered as unrooted.

3.3 Simulation Results

We performed simulations using various sets of parameters. They all give similar results. We present here those obtained with an ancestral sequence of 50,000 base pairs, an average number of indels (*Ind*) in the tree equal to 0, 5 or 10, an average number of reversals (*Rev*) for each leaf ranging from 0 to 4 and two different

substitution rates (*Sub*), 25% and 50% of positions. Each random tree has 16 leaves, contains 13 internal edges and 1,820 quadruplets. The length of the terminal sequence range from 30,000 to 70,000 nucleotides. The average values of RF, NbQ and *MAST* have been evaluated on 200 trials. For the KW distance, value $k = 6$ has been retained, because all the 4,096 words of length 6 are expected in any terminal sequence. But larger value could be used for bacterial genomes around 5 Mb.

Two sets of simulations were performed assuming a constant or variable base composition. Table 1 shows the results when $BC = 0$, bases being equiprobable in any generated sequence. Table 2 corresponds to $BC = 1$. When $Sub = 0.25$ (resp. $Sub = 0.50$) we get sequences having 60% (respectively 75%) of A + T or G + C, as it is often observed among bacterial genomes.

These results clearly show that the MSM distance is more efficient than the three others to recover topology T. For instance the *MAST* and *RF* criteria are generally around 1, which means that only one element is badly placed. Among the other distances, the KW distance provide better results than ACS and ZL, when $BC = 1$. In other sets of simulations, we have tested the ACS distance on random sequences and we observed that the distance values are lower between sequences with the same nucleotide composition than between sequences having a large difference between the A + T and G + C rates. This indicates that the ACS distance tends to join sequences having similar base composition. This becomes obvious when $BC = 1$ and proves that the ACS distance is not adapted to prokaryotic genomes.

Table 1 Average values of Robinson-Foulds, quadruplets and MAST criteria, depending on the number of fragment indels, reversals and substitution rate, for the MSM, ACS, KW and ZL distances ($BC = 0$)

BC = 0			MSM			ACS			KW			ZL		
Ind	Rev	Sub	RF	NbQ	MAST	RF	NbQ	MAST	RF	NbQ	MAST	RF	NbQ	MAST
0	0	.25	0.0	0	.0	1.9	107	2.0	1.5	85	1.6	1.8	96	1.8
5	2	.25	0.1	8	.16	2.2	127	2.3	1.5	81	1.6	3.0	151	2.8
10	4	.25	0.2	14	.23	1.9	110	2.0	1.5	87	1.6	2.7	150	2.7
0	0	.50	0.5	28	.53	2.4	145	2.5	2.9	197	2.9	2.5	150	2.6
5	2	.50	0.6	40	.70	2.5	157	2.7	3.1	235	3.4	2.9	177	2.9
10	4	.50	0.9	65	1.1	2.6	168	2.8	3.0	211	3.1	2.9	196	2.9

Table 2 Average values of Robinson-Foulds, quadruplets and MAST criteria, depending on the number of fragment indels, reversals and substitution rate, for the MSM, ACS, KW and ZL distances ($BC = 1$)

BC = 1			MSM			ACS			KW			ZL		
Ind	Rev	Sub	RF	NbQ	MAST	RF	NbQ	MAST	RF	NbQ	MAST	RF	NbQ	MAST
0	0	.25	0.1	4	0.07	2.1	111	2.2	1.6	92	1.6	2.1	121	2.2
5	2	.25	0.2	8	0.20	2.1	137	2.3	1.7	94	1.9	3.4	190	3.2
10	4	.25	0.3	15	0.28	2.1	110	2.2	1.6	102	1.8	3.0	165	3.0
0	0	.50	0.9	54	.98	7.2	659	5.9	3.6	256	3.6	5.4	433	4.8
5	2	.50	1.0	61	1.0	7.5	670	6.0	3.7	274	3.8	6.9	581	5.8
10	4	.50	1.2	71	1.2	7.7	697	6.1	3.9	275	4.1	6.6	596	5.5

The ZL distance is always intermediate between ACS and KW even when $BC = 1$ and KW becomes much more accurate for phylogeny than the ACS distance.

4 Conclusion

We described the MSM distance between complete DNA genome sequences for phylogenetic reconstruction, avoiding difficulties arising from orthology recognition and gene alignment. It can be computed in time and memory space proportional to genome length. Simulated data showed that the MSM distance over performs the three other alignment free tested distances, and it is not sensitive to bias in base composition.

The superiority of the MSM distance is essentially due to the fact that it only takes into account *significant* matches having a *minimum* length which strongly varies according to the base composition; it is much higher for two genomes sharing similar G + C rate. Therefore it permits to avoid spurious matches and also spurious grouping of taxa.

The MSM phylogenies from to real genomic data will be described in another paper. One can indicate that the topologies are mainly congruent with references phylogenies based on SSU and LSU rRNA sequences, proving that this distance could be used to study relationship within phylum and is very efficient within families and orders.

References

Amir, A., & Keselman, D. (1997). Maximum agreement subtree in a set of evolutionary trees: metric and efficient algorithms. *SIAM Journal on Computing, 26*, 1656–1669.

Estabrook, G.F. et al. (1985). Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Systematic Zoology, 34*, 193–200.

Guyon F., & Guénoche A. (2010). An evolutionary distance based on maximal unique matches. *Communications in Statistics, 39*(3), 385–397.

Karlin, S., & Burge, C. (1995). Dinucleotide relative abundance extremes: A genomic signature. *Trends in Genetics, 11*, 283–290.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution, 16*, 111–120.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology, 5*, R12.

Otu, H. H., & Sayood, K. (2003). A new sequence distance measure for phylogenetic tree construction. *Bioinformatics, 19*(16), 2122–2130.

Qi, J., Wang, B., & Hao, B. I. (2004). Whole proteome prokaryote phylogeny without sequence alignment: A K-string composition approach. *Journal of Molecular Evolution, 58*(1), 1–11.

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences, 53*, 131–147.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution, 4*, 406–425.

Snel, B., Huynen, M. A., & Dutilh, B. E. (2005). Genome trees and the nature of genome evolution. *Annual Review of Microbiology, 59*, 191–209.

Ulitsky, I., Burnstein, D., Tuller, T., & Chor, B. (2006). The average common substring approach to phylogenomic reconstruction. *Journal of Computational Biology, 13*, 336–350.

Ziv, J., & Lempel, A. (1977). A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory, 23*, 337–343.

Data Quality Dependent Decision Making in Pattern Classification

Josef Kittler and Norman Poh

Abstract Sensory information acquired by pattern recognition systems is invariably subject to environmental and sensing conditions, which may change over time. This may have a significant negative impact on the performance of pattern recognition algorithms. In the past, these problems have been tackled by building in invariance to the various changes, by adaptation and by multiple expert systems. More recently, the possibility of enhancing the pattern classification system robustness by using auxiliary information has been explored. In particular, by measuring the extent of degradation, the resulting sensory data *quality* information can be used with advantage to combat the effect of the degradation phenomena. This can be achieved by using the auxiliary quality information as features in the fusion stage of a multiple classifier system which uses the discriminant function values from the first stage as inputs. Data quality can be measured directly from the sensory data. Different architectures have been suggested for decision making using quality information. Examples of these architectures are presented and their relative merits discussed. The problems and benefits associated with the use of auxiliary information in sensory data analysis are illustrated on the problem of personal identity verification in biometrics.

1 Introduction

Many problems in data analysis are specified by the objectives of data analysis and a set of representative data. The solutions found are then used to process new data acquired in subsequent studies or when the solution is deployed in a future operation. A typical example are classification problems, both supervised and non-supervised, where the available data is used to partition the observation space and once the partition is determined, it is then used to classify new data samples.

J. Kittler (✉)
Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, UK
e-mail: J/Kittler@surrey.ac.uk

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_3,

One of the challenges faced in the subsequent use of the data analysis solutions is a data drift. By drift we understand a transformation of the class populations due to factors that influence the data acquisition. Depending on the nature of the data, these factors relate to changes in the sensor characteristics (e.g. for visual sensors either optics or electronics), the environmental conditions (illumination, background noise, clutter, atmospheric conditions), or the behaviour of the imaged object during data acquisition (pose, motion, deformation, biological evolution such as aging).

A good solution to a classification problem will include some form of protection to any anticipated data changes. There are a number of measures that can be adopted to alleviate the problem of a data drift. For instance, one can enhance the robustness of the solution by using invariant measurements. An alternative is to collect data in all the possible conditions that may be envisaged during the system operation. This approach ensures that the data that serves as a basis for solving the classification problem remains representative in future operation. This is somewhat difficult to accomplish, as, at the outset, it is not always possible to predict all the types of changes that might cause a future drift of the class populations.

Although each observation is a single point in the representation space, it retains its physical meaning defined by the sensor and the sensed object. For instance for an imaging sensor, a point may be an image of an object to be classified. Recalling this true nature of the data, one of the open options is to correct for any drift by means of normalisation. For each factor causing a data drift, a suitable normalisation procedure would have to be applied. For example, if the data drift is induced by illumination variation, a photometric normalisation will be required to compensate for such environmental changes. Similarly, an appropriate normalisation would be needed by other types of degradations. Various normalisation procedures proposed in the literature have been shown to be very effective in stabilising the data, e.g., Poh et al. (2009).

Another effective approach to dealing with a data drift is to use multiple experts. In other words, using a set of solutions to the same classification problem instead of the individually best solution. It is well known that, if these experts provide diverse opinions about the points to be classified, the classification accuracy of the solution is improved. It is less well known that multiple classifier systems also improve the robustness of the solution to data drift. In this paper we pursue this particular approach to the data drift problem. We show that the effectiveness of the multiple expert approach can be enhanced by making the use of information about the data quality. By data quality we mean an objective measure of the data departure from its nominal characteristic. As already indicated, data drift is caused by various factors which will be reflected in the properties of the sensory signals. One can view these signal changes as changes in signal quality. In the normalisation approach discussed earlier we attempt to reverse the signal changes by the application of preprocessing algorithms that aim to stabilise the data to be classified. In the multiple classifier system approach the idea is to express signal changes in the form of quality measures. These quality measures can then be used as auxilary features in the multiple classifier system fusion. As a result, the fused system decision is influenced not only by the expert opinions regarding the respective hypotheses, but also by measures of the

signal quality. We formulate the problem of multiple expert decision making which incorporates quality information. We then show that the use of this auxiliary information leads to further improvement of the system performance under data drift. This is illustrated on data relating to personal identity authentication using facial biometric.

The paper is organised as follows. In the next section we develop a theoretical framework for the quality based fusion of multiple classifier systems. In the formulation adopted the quality information is used as additional features. Accordingly, the decision making in the fusion stage of the resulting multiple classifier system is realised in an augmented feature space. In Sect. 3 we demonstrate this approach on a two class problem of face verification, where the data drift is caused by illumination changes. We show that the use of multiple experts results in performance gains over the best performing expert. These gains are further enhanced by incorporating quality information in the fusion process. The paper is drawn to conclusion in Sect. 4.

2 Theoretical Framework

2.1 Problem Formulation

In order to facilitate the discussion, we shall first present some notation.

- $k \in [1, \ldots, K]$ is the class label (or object types), and there are K classes.
- $\mathbf{y} = F(\mathbf{x})$ is a vector output of a classifier F of K dimensions, typically estimating the posterior probabilities of all K class membership given $\mathbf{x}$, $\mathbf{y} \equiv [y_1, y_2, \ldots, y_K]' = [P(k = 1|\mathbf{x}), P(k = 2|\mathbf{x}), \ldots, P(k = K|\mathbf{x})]'$, where "$'$" denotes the matrix transpose operator. The classifier will assign $\mathbf{x}$ to the class

$$k_* = \max_k P(k|\mathbf{x}). \tag{1}$$

  An example of classifier F giving $\mathbf{y}$ as posterior probabilities of class membership is a neural network with the *softmax* activation function at its output layer (Bishop 1999). Although our interpretation of $\mathbf{y}$ is probabilistic, it is not restricted to this. Other classifier architectures, e.g., support vector machines (SVM) and classifiers outputting ratios of two competing hypotheses (e.g., discriminant functions) can also be used as F, as long as the decision rule in (1) (with maximum or minimum) is applicable. In all cases, the score vector $\mathbf{y}$ can be seen as a *measurement* in the class hypothesis space.
- $\mathbf{q} \in \mathbb{R}^L$ is a vector of quality measures output by L quality detectors. A quality detector is an algorithm designed to assess the quality of the signal from which pattern vector $\mathbf{x}$ originates. If the signal is an image, the measures may include, e.g., resolution, the number of bits per pixel, contrast and brightness as defined by the MPEG standards. These measurements aim directly to measure the quality

of the acquired signal. More examples will be given in Sect. 2.3. In general, these measures will be closely linked to the classifier, F, and will have to be designed with the classifier in mind.

There are several points to note. First, in the presence of data-drift affecting $\mathbf{x}$, the measurement $\mathbf{y}$ will also be affected. This will be manifest in an increase in the *entropy* of $\{y_k|k = 1, \ldots, K\}$, defined as

$$\text{entropy}(\mathbf{y}) = \mathbb{E}_k[\log y_k],$$

where $\{y_k|k = 1, \ldots, K\}$ are elements in $\mathbf{y}$.

Second, the hypothesis space $\mathbf{y}$ does not need to reflect probabilities. For instance, one can apply the *logit transform*, or its generalized version, to each of the elements in $\mathbf{y}$. This is a one-to-one *order preserving* transformation. The generalized logit transform is defined as (Dass et al. 2006):

$$\hat{y}_k = \log\left(\frac{y_k - a}{b - y_k}\right) \tag{2}$$

for an output variable y_k bounded in $[a, b]$. An important advantage of this transformation is that the processed vector $\hat{\mathbf{y}}$ appears to be much more normally distributed, rather than skewed as in the probability space. This can significantly improve the design of a fusion classifier in any subsequent stage of decision making (stacked generalizer). An example of this transformation is shown in Fig. 1.

In the multiple expert paradigm, one would construct multiple estimates of $\mathbf{y}$. Let i be the index of the i-th expert and let there be $i = 1, \ldots, N$ experts. We shall introduce $\mathbf{y}_i = F_i(\mathbf{x})$ as the output of expert F_i (each observing the same data sample, $\mathbf{x}$). At this point, it is also convenient to define $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]'$ as a concatenation of the outputs of all N experts. Note that an element of $\mathbf{Y}$, Y_{ik}, indicates the output of the i-th expert for the k-th class hypothesis. In order to integrate the

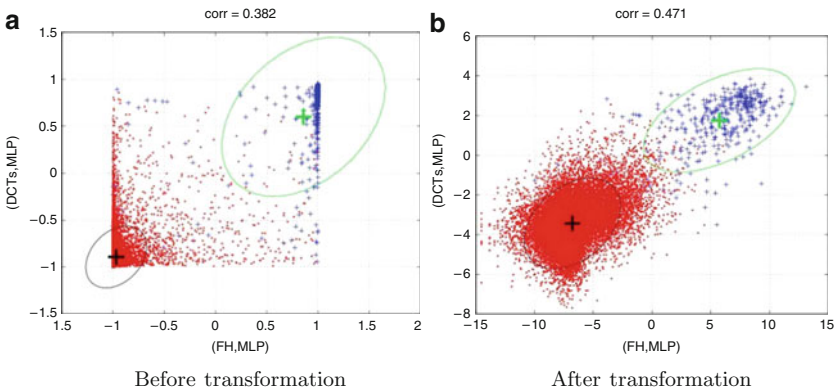


Fig. 1 The effect of applying the generalized logit transformation to two expert outputs

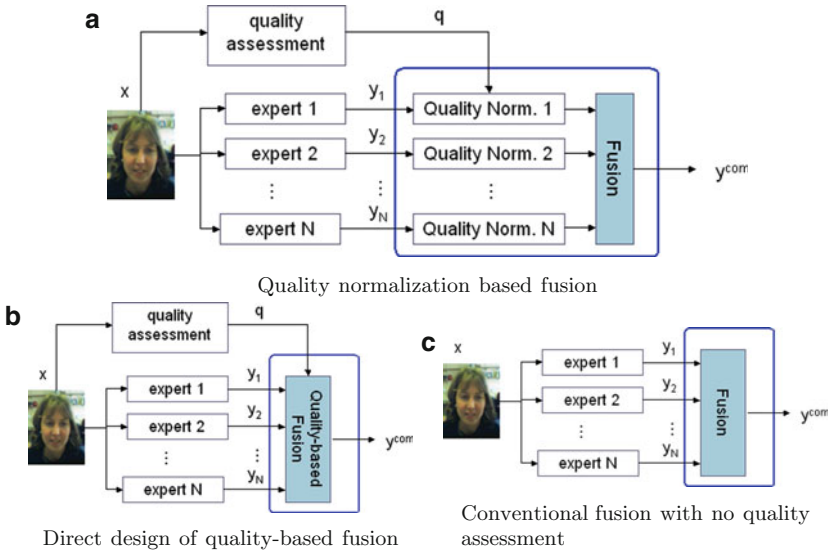


Fig. 2 Two possible architectures for implementing quality-based fusion

opinions of the respective experts, we need to design a fusion mechanism capable of handling the output of multiple experts along with the quality measures:

$$G : \mathbf{Y}, \mathbf{q} \rightarrow \mathbf{y}^{com} \tag{3}$$

where $\mathbf{y}^{com} \in \mathbb{R}^K$. G is also known as *quality-based* fusion.

The function G estimates the posterior probability of the respective hypotheses, where the element in $\mathbf{y}^{com}$, y_k^{com}, is estimated by:

$$\mathbf{y}_k^{com} = P(k|\mathbf{Y}, \mathbf{q}) \tag{4}$$

Two possible architectures for realising G are shown in Fig. 2. In the first approach (Fig. 2a), the quality measures are used to normalise each individual expert output $\mathbf{y}_i$:

$$\mathbf{y}_i^{norm} = G_i(\mathbf{y}_i, \mathbf{q}), \text{ for all } i \tag{5}$$

The normalized outputs are then combined using another classifier or a fixed rule (sum or product) in order to obtain $\mathbf{y}^{com}$. A typical example of the latter case is the sum fusion

$$\mathbf{y}^{com} = \sum_i \mathbf{y}_i^{norm}$$

The second approach (see Fig. 2b) solves (3) directly. In the absence of the quality measures, both approaches reduce to the conventional fusion (i.e. without quality measures), shown in Fig. 2c. It will shown in Sect. 3 that under changing test

conditions, the quality-based fusion systematically outperforms the conventional one (Kittler et al. 2007).

The first approach, combined with a fixed rule, offers an attractive modular solution in the sense that each expert output is processed individually. Effectively, the dimension of the quality augmented hypothesis space is $K+L$ (the sum of length of the vectors $\mathbf{y}_i$ and $\mathbf{q}$). This is significantly smaller than that of the second approach, which is $K \times N + L$ (the sum of dimensions of $\mathbf{Y}$ and $\mathbf{q}$), or that of the first approach with a general fusion scheme where the dimensionality is $K \times N$.

The increased dimensionality in the second approach can be seen as a weakness because many more parameters need to be estimated at the same time. This weakness is, however, outweighed by an important advantage: the ability to handle the dependency among the system outputs. Since each expert observes the same input sample $\mathbf{x}$, their outputs are necessarily dependent. The first approach with a general fusion scheme can still handle any output dependency in the fusion module. However, the first approach with a fixed rule, such as the Naive Bayes, ignores the expert output dependency and this may be reflected in degraded performance.

2.2 Quality-Based Fusion

This section aims to estimate (4), bearing in mind that any classifier, even those that do not output probabilities (e.g., SVM) can be used after an appropriate normalisation.

We shall explore two different approaches: generative and discriminative.

In order to facilitate the discussion for the generative approaches, we shall use graphical models (Bishop 2007), also known as Bayesian networks (Jensen 1996). A graphical model is a graph with directed arrows representing conditional probabilities. A node in the graph is a variable. An arrow from variable A to variable B specifies their causal relationship, i.e., the conditional probability of B given A, i.e., $p(B|A)$.

Two possible graphical models for modeling the relationship between $\mathbf{y}$ (noting that the index for each expert i is not used here for simplicity) and $\mathbf{q}$ are shown in Fig. 3. The first model (Nandakumar et al. 2008), as depicted in Fig. 3a, attempts to characterize the following joint density

$$p(\mathbf{y}, \mathbf{q}, k) = p(\mathbf{y}, \mathbf{q}|k)P(k) \tag{6}$$

whereas the second model (Fig. 3b) achieves this slightly differently:

$$p(\mathbf{y}, \mathbf{q}, k) = p(\mathbf{y}|k, \mathbf{q})p(\mathbf{q}|k)P(k), \tag{7}$$

Note that the second model involves the density of quality measures conditioned on class label k. However, as the signal quality is *independent* of the class label k, i.e., the quality measures cannot be used to distinguish among different classes of

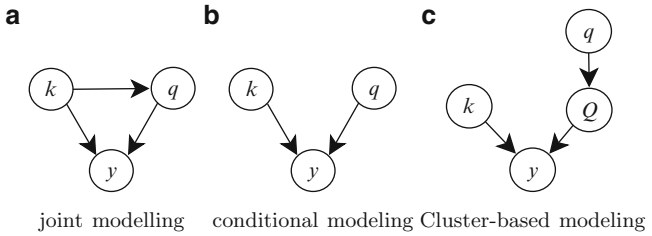


Fig. 3 Three graphical models that capture the relationship between match scores y, quality measures q an the class label k in different ways

objects, modelling $p(\mathbf{q}|k)$ is unnecessary. More importantly, an impending problem with the implementation of the second model is the need to estimate the conditional density $p(\mathbf{y}|k,\mathbf{q})$. Because the conditioning variable $\mathbf{q}$ is multivariate and continuous, one has to use a multivariate regression algorithm. In comparison, the first approach needs only to estimate $p(\mathbf{y},\mathbf{q}|k)$, a problem which is well understood, since the conditioning variable in this case is discrete.

Although the first approach is preferable, there is an alternative. The third approach (Poh et al. 2007), attempts to model q by introducing a latent discrete variable $Q \in [1,\ldots,\mathtt{Q}]$ (see Fig. 3c). The idea is first to classify q into $\mathtt{Q}$ discrete clusters. Then, the modeling of $p(\mathbf{y}|\mathbf{q},k)$ can be achieved as follows:

$$p(\mathbf{y}|\mathbf{q},k) = \sum_{Q} p(\mathbf{y},Q|\mathbf{q},k) = \sum_{Q} p(\mathbf{y}|Q,k)P(Q|\mathbf{q})$$

where $P(Q|\mathbf{q})$ is the posterior probability of cluster Q given the quality measures $\mathbf{q}$. The cluster-based approach can effectively provide a simpler method of implementing the second approach.

If one uses a mixture of Gaussian (Bishop 1999) as a clustering algorithm, estimating

$$p(\mathbf{q}) = \sum_{Q} p(\mathbf{q}|Q)P(Q),$$

then the posterior probability $p(Q|\mathbf{q})$ can be estimated via the Bayes rule:

$$P(Q|\mathbf{q}) = \frac{p(\mathbf{q}|Q)P(Q)}{p(\mathbf{q})}.$$

Once the density $p(\mathbf{y},\mathbf{q}|k)$ (for the first approach) or $p(\mathbf{y}|\mathbf{q},k)$ (for the second and third approach) is estimated, for the generative model, the posterior probability of class membership can be obtained by using the Bayes rule:

$$P(k|\mathbf{y},\mathbf{q}) = \frac{p(\mathbf{y},\mathbf{q}|k)P(k)}{\sum_{k'} p(\mathbf{y},\mathbf{q}|k')P(k')} \tag{8}$$

The second and third model can be estimated similarly by replacing the term $p(\mathbf{y}, \mathbf{q}|k)$ with $p(\mathbf{y}|\mathbf{q}, k)$.

Extending this concept to estimating $P(k|\mathbf{Y}, \mathbf{q})$, in the context of multiple classifiers, is straightforward. Following the discussion in Sect. 2.1, one can employ the normalization-based strategy or the joint modeling strategy. For the first strategy, one can employ the sume rule, or the Naive Bayes principle, i.e., by estimating $p(\mathbf{Y}, \mathbf{q}|k)$ as $p(\mathbf{Y}, \mathbf{q}|k) = \prod_i p(\mathbf{y}_i, \mathbf{q}|k)$, and then applying the Bayes rule in order to obtain $P(k|\mathbf{Y}, \mathbf{q})$. For the second strategy, one simply replaces every appearance of $\mathbf{y}$ by $\mathbf{Y}$ in this section.

As an alternative to the generative approach, one can estimate $P(k|\mathbf{y}, \mathbf{q})$ directly. For instance, the posterior probability $P(k|\mathbf{y}, \mathbf{q})$ (for all K classes) can be implemented using a multilayer perception with a softmax output layer (Bishop 1999). Alternatively, for non-probabilistic classifiers, one can use a multi-class SVM, to learn discriminative functions $H_k(\mathbf{Y}, \mathbf{q})$. To obtain the posterior probabilities of class membership k, we use

$$P(k|\mathbf{Y}, \mathbf{q}) = \frac{\exp\{H_k(\mathbf{Y}, \mathbf{q})\}}{\sum_j \exp\{H_j(\mathbf{Y}, \mathbf{q})\}} \tag{9}$$

2.3 Data Quality Assessment

The quality of sensory signal giving rise to data $\mathbf{x}$, respectively $\mathbf{y}$ is multifaceted and cannot be captured by a single quality measure. Instead, a collection of quality measures, $q_1, \ldots, q_P$ should be computed, as implied in the formulation given in Sect. 2.1. For instance, there are many measures that have been proposed to characterize image quality. These include focus, resolution, image size, uniformity of illuminations, background noise, object pose, etc. If all these measures are treated as separate features augmenting the dimensionality of the score space, then its size may grow disproportionately, potentially leading to over training problems. The likelihood of poor generalization is high, especially in view of the fact that quality measures themselves do not convey discriminatory information. This may further be aggravated by the small sample size problems plaguing some classification applications.

Second, signal quality is not an absolute concept. Suppose a classification system is designed using a set of training data acquired with a web camera, but an operational test is conducted using images captured with a camera of much better quality. From the point of view of the classification algorithm, the better quality image will actually appear as degradation. Thus a quality assessment should be carried out in the context of a reference defined by the system design conditions. In fact the situation is even worse, as the existing approaches to quality based fusion do not take into account the properties of the classification algorithms. For instance, if one system uses an algorithm that can correct for illumination or object pose problems, then the supplied image quality information will be misleading and may affect the

performance of the system adversely. On the other hand, for algorithms that cannot compensate for changes in illumination and pose the quality information is likely to be crucial. Thus signal quality assessment cannot be algorithm independent either. This raises a fundamental question how signal quality should be defined, whether it can be measured directly from the sensory data, or whether it should be derived by the classification algorithm itself using the internal knowledge about its capabilities.

3 An Illustration of the Benefits of the Quality Based Fusion

We illustrate the benefit of quality based fusion in the presence of data drift on the problem of identity verification using face biometrics. The experiments are conducted on the XM2VTS database (Matas et al. 2000) and its degraded section (Messer et al. 2006). The problem is illustrative of data drift caused by the changing conditions of the acquisiton environment and hence the quality of the data. The original database which was used for training the class models contains mugshot images with well controlled illumination. The darkened section contains images taken under strong side illumination, which is known to degrade significantly face verification performance (Messer et al. 2006). Examples of these images are shown in Fig. 4.

The database contains 295 subjects, which includes 200 subjects selected to be clients, 25 to be impostors for the algorithm development (training), 70 to be impostors for algorithm evaluation (testing). For each subject, the face images are acquired in four sessions; the first three are used for training the classifiers and the last one for testing. We consider the dark dataset with left illumination as the "fifth session" and the one with right illumination as the "sixth" session.

We used a set of proprietary quality measures developed by Omniperception Ltd for the face image quality assessment. These measures are: "frontal quality", measuring the deviation from the frontal face; and "illumination quality", quantifying the uniformity of illumination of the face. It should be noted that none of these quality detectors were designed specifically to distinguish the three strong dominant quality states of the face images in the XM2VTS database: good illumination,

Fig. 4 Frontal and side illumination of a subject taken from the XM2VTS database

left illumination and right illumination. Using the above quality measures makes the problem of quality-dependent fusion more challenging.

The classifiers used for the face experts in this paper can be found in Heusch et al. (2006). There are two classifiers with three types of pre-processing, hence resulting in a matrix of six classifiers. The two classifiers used are Linear Discriminant Analysis (LDA) with correlation as a measure of similarity (Kittler et al. 2000) and Gaussian Mixture Model (GMM) with maximum a posteriori adaptation, described in Reynolds et al. (2000). The use of the GMM in face authentication can be found in Cardinaux et al. (2006). The face pre-processing algorithms used include the photometric normalisation as proposed by Gross and Brajovic (2003), histogram equalisation and local binary pattern (LBP) as reported in Heusch et al. (2006). The feature extraction and classification algorithms are implemented using the open-source Torch Vision Library.[1]

With the availability of six face experts, we performed exhaustive fusion, each time combining 2, 3, etc., experts until all six are used. This results in 63 combinations. The two architectures, namely quality normalization-based and joint fusion approaches (shown in Fig. 2a and b) are then compared with the baseline system in Fig. 2c. Since this is a binary classification problem (a person is either a genuine user or an impostor), logistic regression was used to approximate the posterior probabilities of class membership in all cases. The results of the 63 fusion tasks are shown in Fig. 5.

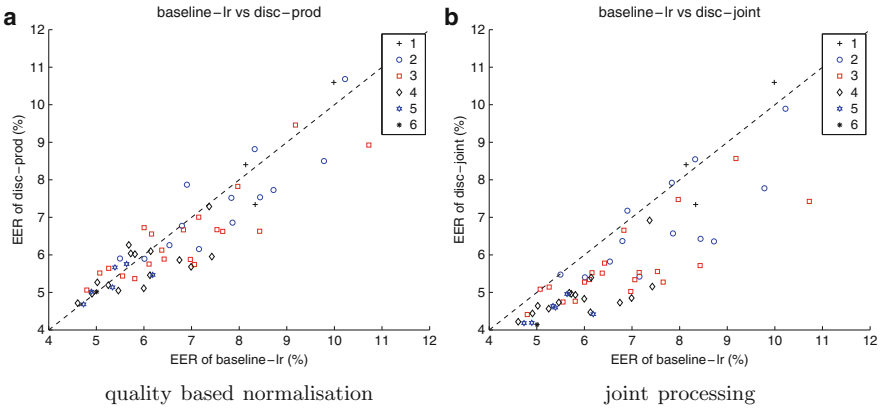


Fig. 5 Comparison of (a) quality based normalisation (Architecture 1) and (b) joint processing (Architecture 2); in the Y-axes with respect to the baseline system without using quality (Architecture-3); in the X-axes. Each point in the figures are the *a posteriori* EER (%) of one of the possible 63 fusion tasks. In both figures, the numbers in the legend are the number of experts used in one of the 63 fusion tasks

[1] Available at "http://torch3vision.idiap.ch". See also a tutorial at "http://www.idiap.ch/marcel/labs/faceverif.php".

For performance evaluation, we used Equal Error Rate (EER). This error is often used in the face recognition community to handle the case of highly unbalanced class priors. EER is defined as the average of False Acceptance Rate (FAR) and False rejection Rate (FRR). FAR is also known as *false alarm rate* whereas FRR is also known as *miss detection rate*.

As can be observed, the proposed approach using $\{\mathbf{Y}, \mathbf{q}\}$ is almost always better than the baseline fusion approach using only $\mathbf{Y}$. The average observed relative improvement in the system robustness to data drift is about 25% but up to 40% can be attained.

4 Conclusions

A data drift caused by changes of the environment, sensor characteristics and object representation in sensory data acquisition can seriously degrade the performance of pattern classification systems. We proposed a solution which is based on the protective measures against data drift offered by the paradigm of multiple classifier systems. We showed that by incorporating sensory data quality information in the fusion stage of the multiple classifier system, considerable robustness to data drift can be achieved. A framework for quality assisted fusion of multiple experts has been developed and its variants discussed. The proposed approach has been demonstrated on the problem of personal identity verification using facial biometrics. The improvement in handling a data drift caused by illumination changes in face image acquisition was 25% on average and could be as high as 40%.

Acknowledgements This work was supported partially by the advanced researcher fellowship PA0022 121477 of the Swiss National Science Foundation and by the EU-funded Mobio project (www.mobioproject.org) grant IST-214324.

References

Bishop, C. (1999). *Neural networks for pattern recognition*. Oxford: Oxford University Press.

Bishop, C. M. (2007). *Pattern recognition and machine learning*. Berlin, Heidelberg, New York: Springer.

Cardinaux, F., Sanderson, C., & Bengio, S. (2006). User authentication via adapted statistical models of face images. In *IEEE Transactions on Signal Processing* (pp. 361–373).

Dass, S. C., Zhu, Y., & Jain, A. K. (2006). Validating a biometric authentication system: Sample size requirements. *IEEE Transactions on Pattern Analysis and Machine, 28*(12), 1902–1319.

Gross, R., & Brajovic, V. (2003). An image preprocessing algorithm for illumination invariant face recognition. In *4th International Conference on Audio and Video-Based Biometric Person Authentication (AVBPA'03)* (pp. 10–18).

Heusch, G., Rodriguez, Y., & Marcel, S. (2006). Local binary pattern as an image preprocessing face authentication. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06)* (pp. 9–14). Washington, DC.

Jensen, F. V. (1996). *An introduction to Bayesian networks*. Berlin, Heidelberg, New York: Springer.

Kittler, J., Li, Y., & Matas, J. (2000). On matching score for LDA-based face verification. In *British Machine Vision Conference (BMVC)*. University of Surrey, UK.

Kittler, J., Poh, N., Fatukasi, O., Messer, K., Kryszczuk, K., Richiardi, J., & Drygajlo, A. (2007). Quality dependent fusion of intramodal and multimodal biometric experts. In *Proceedings of SPIE Defense and Security Symposium, Workshop on Biometric Technology for Human Identification* (Vol. 6539).

Matas, J., Hamouz, M., Jonsson, K., Kittler, J., Li, Y., Kotropoulos, C., Tefas, A., Pitas, I., Tan, T., Yan, H., Smeraldi, F., Begun, J., Capdevielle, N., Gerstner, W., Ben-Yacoub, S., Abdeljaoued, Y., & Mayoraz, E. (2000). Comparison of face verification results on the XM2VTS database. In *Proceedings of the 15th International Conference on Pattern Recognition* (Vol. 4, pp. 858–863). Barcelona.

Messer, K., Kittler, J., Short, J., Heusch, G., Cardinaux, F., Marcel, S., Rodriguez, Y., Shan, S., Su, Y., & Gao, W. (2006). Performance characterisation of face recognition algorithms and their sensitivity to severe illumination changes. In *LNCS 3832, Proceedings of the International Conference on Biometrics (ICB'06)* (pp. 1–11). Hong Kong.

Nandakumar, K., Chen, Y., Dass, S. C., & Jain, A. K. (2008). Likelihood ratio based biometric score fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (Vol. 30, pp. 342–347).

Poh, N., Heusch, G., & Kittler, J. (2007). On combination of face authentication experts by a mixture of quality dependent fusion classifiers. In *LNCS 4472, Multiple Classifiers System (MCS)* (pp. 344–356). Prague.

Poh, N., Bourlai, T., & Kittler, J. (2009). Quality-based score normalisation with device qualitative information for multimodal biometric fusion. *IEEE Transactions on Systems, Man, and Cybernatics (part B)* (in press).

Reynolds, D. A., Quatieri, T., & Dunn, T. (2000). Speaker verification using adapted guassian mixture models. In *Digital Signal Processing* (pp. 19–41).

Clustering Proteins and Reconstructing Evolutionary Events

Boris Mirkin

Abstract The issue of clustering proteins into homologous families has attracted considerable attention by researchers. On one side, many databases of protein families have been developed by using relatively simple clustering methods and a lot of manual curation. On the other side, more elaborated clustering approaches have been used, yet with a very limited degree of success. This paper advocates an approach to clustering protein families involving the knowledge of protein functions to adjust the parameter of similarity scale shift. We proceed to reconstruct HPF evolutionary histories to both further narrow down the choice of the cluster solution and interpret clusters.

1 Introduction: Clustering and Knowledge Feedback

Clustering is conventionally applied for deriving protein families (see, for example, Thompson et al. 1994; Tatusov et al. 2000; Bader and Hogue 2003; Kawaji et al. 2004; Chen et al. 2006; Mirkin et al. 2006; Paccanaro et al. 2006; Brown et al. 2007; Poptsova and Gogarten 2007; Mirkin et al. 2010).

Our clustering method falls within the so-called data recovery approach applied to the similarity data. According to this approach similarity data are considered as weighted sums of "ideal" structures such as partitions or clusters, the clusters and their intensity weights being determined by minimizing the differences between the given similarity data and those generated by the putative model. We extract clusters one-by-one (Mirkin 1987, 1996), to both facilitate the search and supply

B. Mirkin
School of Computer Science, Birkbeck University of London, Malet Street, London, WC1 7HX, UK
e-mail: mirkin@dcs.bbk.ac.uk
and
Department of Applied Mathematics, Higher School of Economics, Kirpichnaya 33/5, Moscow, RF
e-mail: bmirkin@yandex.ru

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_4,

meaningful estimates of their intensity and contribution to the data scatter. In a data recovery clustering model, there is a parameter, analogous to the intercept of the regression line, that plays the role of a prior similarity shift. This parameter also is a 'soft' similarity threshold, so that entities whose similarity is less than its value are unlikely to get combined in the same cluster. The parameter's value may strongly affect the number and contents of the clusters, and it can be derived according to the least-squares criterion. However, as we shall illustrate, a better choice may be made by using proteomics knowledge.

A clustering method, derived from the data recovery approach, can be applied to evolutionary intergenomic studies in which clusters are interpreted as homologous protein families (HPFs). The proteins in each of these families are assumed to be inherited from the same ancestor, so that an HPF can be parsimoniously mapped to an evolutionary tree on the set of genomes under consideration, thereby reconstructing the HPF's evolutionary history. Obviously, the reconstructed histories may critically depend on the level of aggregation: a highly aggregated family intersecting all or almost all genomes would be mapped to the last common ancestor. However, if the family is partitioned, the parts would be mapped to different, more recent, ancestors. These two mappings would lead to two different histories of the function of the HPF under consideration. As the level of aggregation of proteins depends on the value of the similarity threshold/shift, the evolutionary mapping of protein families can be used for fine tuning that value by analyzing the consistency between the reconstructed histories and other data available.

To determine an appropriate value for the similarity shift, we analyze a set of pairs of HPFs whose functions are known. The expectation is that proteins with the same function should be more similar to each other than would be proteins with dissimilar functions. This should indicate an appropriate similarity value that could distinguish those pairs that should be in the same cluster from those that should not. The actual distribution of similarity scores may turn out to be more complex than we would hope, so that not one but two reasonable similarity shift values emerged: one would guarantee that HPFs with dissimilar functions would be in different clusters, whereas the other would give the minimum error in separating protein pairs with similar and dissimilar functions. The final choice, however, requires further knowledge of the genomes, viz. the consistency between the suggested ancestral reconstructions and gene arrangements.

Therefore, our approach involves two phases of interference of the clustering and proteome knowledge: one, passive, takes in the knowledge to adjust the values of a clustering parameter, and the second, active, makes use of cluster-based evolutionary histories of protein functions.

The rest of the paper is organized as follows. Section 2 describes the data recovery approach to clustering similarity data. Section 3 is devoted to a description of the results of clustering protein families by using the knowledge of protein functions to identify similarity shift values. Section 4 describes some results involving the reconstructed evolutionary histories. In Section 5 we conclude and outline possible future work.

2 Clustering Using the Data Recovery Approach

2.1 Additive Clustering and One-by-One Iterative Extraction

Let I be a set of entities under consideration and let $A = (a_{ij})$ be a symmetric matrix of similarities (or, synonymously, proximities or interactions) between entities $i, j \in I$.

The additive clustering model (Shepard and Arabie 1979; Mirkin 1976, 1987) assumes that the similarities in A are generated by a set of 'additive clusters' $S^k \subseteq I$, $k = 0, 1, \ldots, K$, in such a way that each a_{ij} approximates the sum of the intensities of those clusters that contain both i and j:

$$a_{ij} = \sum_{k=1}^{K} \lambda_k s_i^k s_j^k + \lambda_0 + e_{ij}, \tag{1}$$

where $s^k = (s_i^k)$ are the membership vectors of the unknown clusters S^k and λ_k are their intensities, $k = 1, 2, \ldots, K$; e_{ij} are the residuals to be minimised.

The intercept value λ_0 can be interpreted as the intensity of the universal cluster $S_0 = I$ that must be part of the solution and, on the other hand, it has a meaning of the similarity shift, with the shifted similarity matrix $A' = (a'_{ij})$ defined by $a'_{ij} = a_{ij} - \lambda_0$. Equation (1) for the shifted model can be rewritten in an obvious way so that it expresses a'_{ij} through clusters $k = 1, \ldots, K$ by moving λ_0 onto the left. The role of the intercept λ_0 in (1) as a 'soft' similarity threshold is of special interest when λ_0 is user specified because the shifted similarity matrix a'_{ij} may lead to different clusters at different λ_0 values.

To fit model (1), we apply one-by-one cluster extracting strategy by minimizing, at each step $k = 1, \ldots, K$ criterion

$$L^2(S, \lambda) = \sum_{i,j \in I} (a'_{ij} - \lambda s_i s_j)^2 \tag{2}$$

and setting the found solutions S and λ as S_k and λ_k, respectively. Obviously, the optimal λ_k is the average of residual similarities a'_{ij} within S_k. The residual similarities a'_{ij} are updated after each step k by subtracting $\lambda_k s_{ik} s_{jk}$.

This strategy leads to the following decomposition of the data scatter into the contributions of the extracted clusters S^k ("explained" by the model) and the minimized residual square error (the "unexplained" part) (Mirkin 1987):

$$(A', A') = \sum_{k=1}^{K} [s^{kT} A^k s^k / s^{kT} s^k]^2 + (E, E) \tag{3}$$

The inner products (A', A') and (E, E) denote the sums of the squares of the elements of the matrices, considering them as vectors.

2.2 One Cluster Clustering

In this section, we turn to the problem of minimization of (2) for extraction of a single cluster. It should be noted that if A is not symmetric, it can be equivalently changed for symmetric $\hat{A} = (A + A^T)/2$ (Mirkin 1976, 1996). For the sake of simplicity, in this section, we assume that the diagonal entries a_{ii} are all zero.

2.2.1 Pre-specified Intensity

When the intensity λ of the cluster to be found is pre-specified, criterion (2) can be expressed as

$$L^2(S, \lambda) = \sum_{i,j \in I} (a_{ij} - \lambda s_i s_j)^2 = \sum_{i,j \in I} a_{ij}^2 - 2\lambda \sum_{i,j \in I} (a_{ij} - \lambda/2) s_i s_j \quad (4)$$

For $\lambda > 0$, minimizing (4) is equivalent to maximizing the sum on the right,

$$f(S, \pi) = \sum_{i,j \in I} (a_{ij} - \pi) s_i s_j = \sum_{i,j \in S} (a_{ij} - \pi). \quad (5)$$

This implies that, for any entity i to be added to or removed from the S under consideration, the difference between the value of (5) at the resulting set and its value at S, $f(S \pm i, \pi) - f(S, \pi)$, is equal to $\pm 2 f(i, S, \pi)$ where

$$f(i, S, \pi) = \sum_{j \in S} (a_{ij} - \pi) = \sum_{j \in S} a_{ij} - \pi |S|.$$

This gives rise to a local search algorithm for maximizing (5): start with $S = \{i^*, j^*\}$ such that $a_{i^* j^*}$ is maximum element in S, provided that $a_{i^* j^*} > \pi$. An element $i \notin S$ may be added to S if $f(i, S, \pi) > 0$; similarly, an element $i \in S$ may be removed from S if $f(i, S, \pi) < 0$. The greedy procedure ADDI (Mirkin 1987) iteratively finds an $i \notin S$ maximizing $+f(i, S, \pi)$ and an $i \in S$ maximizing $-f(i, S, \pi)$, and takes the i giving the larger value. The iterations stop when this larger value is negative. The resulting S is returned along with its contribution to the data scatter, $4\pi \sum_{i \in S} f(i, S, \pi)$.

To reduce the dependence on the initial S, a version of ADDI can be utilized by starting from singleton $S = \{i\}$, for each $i \in I$, and finally selecting, from all S found at different i, that S that contributes most to the data scatter, i.e. minimizes the square error L^2 (2).

The algorithm CAST (Ben-Dor et al. 1999), popular in bioinformatics, is a version of the ADDI algorithm, in which $f(i, S, \pi)$ is reformulated as $\sum_{j \in S} a_{ij} - \pi |S|$ and $\sum_{j \in S} a_{ij}$ is referred to as the affinity of i to S.

Another property of the criterion is that $f(i, S, \pi) > 0$ if and only if the average similarity between a given $i \in I$ and the elements of S is greater than π, which

means that the final cluster S produced by ADDI/CAST is rather tight: the average similarities between $i \in I$ and S is at least π if $i \in S$ and no greater than π if $i \notin S$ (Mirkin 1987).

Changing the threshold π should lead to corresponding changes in the optimal S: the greater π is, the smaller S will be (Mirkin 1987).

2.2.2 Optimal Intensity

When λ in (4) is not fixed but chosen to further minimize the criterion, it is easy to prove that:

$$L^2(S, \lambda) = (A, A) - [s^T As/s^T s]^2, \tag{6}$$

The proof is based on the fact that the optimal λ is the average similarity $a(S)$ within S, i.e.,

$$\lambda = a(S) = s^T As/[s^T s]^2, \tag{7}$$

since $s^T s = |S|$.

The decomposition (6) implies that the optimal cluster S must maximize the criterion

$$g^2(S) = [s^T As/s^T s]^2 = a^2(S)|S|^2 \tag{8}$$

or its square root, the Raleigh quotient,

$$g(S) = s^T As/s^T s = a(S)|S| \tag{9}$$

over all binary vectors s.

To maximize $g(S)$, one may utilize the ADDI-S algorithm (Mirkin 1987), which is a version of the algorithm ADDI/CAST, described above, in which the threshold π is recalculated after each step as $\pi = a(S)/2$, corresponding to the optimal λ in (7).

A property of the resulting cluster S, similar to that for the constant threshold case, holds: the average similarity between i and S is at least half the within-cluster average similarity $a(S)/2$ if $i \in S$, and at most $a(S)/2$ if $i \notin S$.

ADDI-S utilizes no ad hoc parameters, so the number of clusters is determined by the process of clustering itself. However, changing the similarity shift λ_0 may affect the clustering results, which can be of advantage in contrasting within – and between – cluster similarities.

3 Proteome Knowledge in Determining Similarity Shift

3.1 Protein Families and Evolutionary Tree

The concept of homologous protein family, HPF, can be considered an empirical expression of the concept of gene as a unit of heredity in the intergenomic evolutionary studies (Tatusov et al. 2000; Alba et al. 2001). As such the HPF is an important

instrument in the analysis of the evolutionary history of the function that it bears. The evolutionary history of a set of genomes under consideration is depicted as an evolutionary tree, or phylogeny, whose leaves are one-to-one labelled by genomes of the set, and internal nodes correspond to hypothetical ancestors. An HPF can be mapped to the tree in the following natural way (Mirkin et al. 2003).

First, the HPF is assigned to the leaves corresponding to genomes containing its members. Then the pattern of belongingness can be iteratively extended to all the ancestor nodes in a most parsimonious or most likely way. For example, if each child of a node bears a protein from the HPF then the node itself should bear the same gene itself, because it is highly unlikely that the same gene emerged in the children independently (Mirkin et al. 2003, 2006). Having annotated the evolutionary tree nodes with hypothetical evolutionary histories of various HPFs, realistic conclusions of possible histories and mechanisms of evolution of biomolecular function may be drawn for the purposes of both theoretical research and medical practice.

Assignment of proteins to HPFs is often determined with a large manual component because the degree of similarity between proteins within an alignment of protein sequences, typically, with PSI-BLAST (Altschul et al. 1997) or the like, is not always sufficient to automatically identify the families. This is why a two-stage strategy for identifying HPFs has been considered in (Mirkin et al. 2006). According to this strategy, HPFs are created, first, as groups of proteins that have a common motif, a contingent fragment of protein sequence that is similar in all HPFs members by using a software such as the XDOM (Gouzy et al. 1997; Alba et al. 2001). This motif represents a relatively well conserved segment of the genetic material that can be associated with a protein function. Obviously such motif defined HPFs may be overly fragmented since (1) some functional sites, contiguous in the spatial fold, may correspond to dis-contiguous fragments of protein sequences, and (2) multifunctional proteins may bear resemblances to different proteins at different places.

The fragmented HPFs may lead then to wrong reconstructions of functional histories because if they bear similar proteins and thus should be combined into a single aggregate HPF, then its origin ought to be in the ultimate ancestor corresponding to the tree root rather than in separate subtrees of the phylogeny.

Therefore, the next stage of the strategy is to cluster the first stage HPFs into larger aggregations. Since entities at this stage are not single proteins but protein families, one needs to score similarities between families rather than single proteins, which we do by using the set-theoretic similarity – not between HPFs themselves – but rather between their neighborhoods defined by using PSI-BLAST (Altschul et al. 1997). Given an HPF, this approach works as follows. First, for every protein from the HPF a list of similar proteins is created using PSI-BLAST. Second, these lists are combined into a set of proteins, the neighborhood, according to a majority rule. Third, a set-similarity index values are computed between the HPF neighborhoods. One can notice such advantages of this approach as

- Accuracy of protein alignments because only neighboring proteins are aligned here;

- Better capturing functional properties of the proteins. For example, the glycoprotein H like protein of murine herpesvirus 4 (gi: 1246777) and the UL22 protein of Bovine herpesvirus 1 (gi: 1491636) have minimal sequence identity (15%, identified on the second PSI-BLAST iteration), and have been assigned to separate HPFs within the VIDA database (Alba et al. 2001). However, their sets of homologous protein neighbors (with 20% or greater sequence identity), contain 25 and 20 sequences, respectively, and have 14 common proteins, making the overlap between the homologous protein lists 63% on average.
- The evolutionary timing can be caught up at different majority thresholds (Mirkin et al. 2006) as an alternative to relying on statistical frequency profiles in PSI-BLAST (Altschul et al. 1997).

The idea of employing neighborhoods to measure similarities between entities stems from earlier work, see for example Jarvis and Patrick (1973). Our clustering model leads to the index of average overlap $mbc = (n/n1 + n/n2)$ for scoring the similarity between subsets of sizes $n1$ and $n2$ whose overlap is of size n.

The data for this analysis come from studies of herpesvirus – a pathogene highly affecting both animals and humans. A set of 30 complete herpesvirus genomes covering the so-called α, β and γ herpesvirus superfamilies that differ by the tissue in which the virus resides, have been extracted by authors of Mirkin et al. (2006) from the herpesvirus database VIDA (Alba et al. 2001) and an evolutionary tree has been built over the genomes using the neighbor joining algorithm from PHYLIP package (Felsenstein 2001) (see Fig. 1). This tree totally agrees with the previously published instances of herpesvirus phylogenies (Davison 2002; McGeoch et al. 2006), except for the uncertainty fragments acknowledged in these publications. A set of 740 HPFs represented in these 30 genomes have been extracted from the VIDA database too (Alba et al. 2001).

3.2 Utilizing Knowledge of Proteome

To choose a right λ_0 value in model (1), one should use the external knowledge of the proteome, independent of sequence similarity estimates, for example, of functional activities of the proteins. Each HPF is supposed to have a biomolecular function (for examples of function see Table 1 below), though unfortunately functions of most proteins are unknown yet. We can use those HPFs that have similar functions versus those that are not to choose the 'right' level of the similarity shift. Operationally, we consider proteins as functionally similar if they are consistently named between the herpesvirus genomes and/or they share the same known function. The similarity shift value should be taken such that similarities between dissimilar HPFs get negative after the shift while those between similar HPFs remain positive. To implement this idea, we analyzed 287 available pairs of HPFs with known function and positive similarity value. Among them, no dissimilar pair has a greater *mbc* similarity than 0.66, which should imply that the shift value $\lambda_0 = 0.67$ confers specificity for the production of APFs.

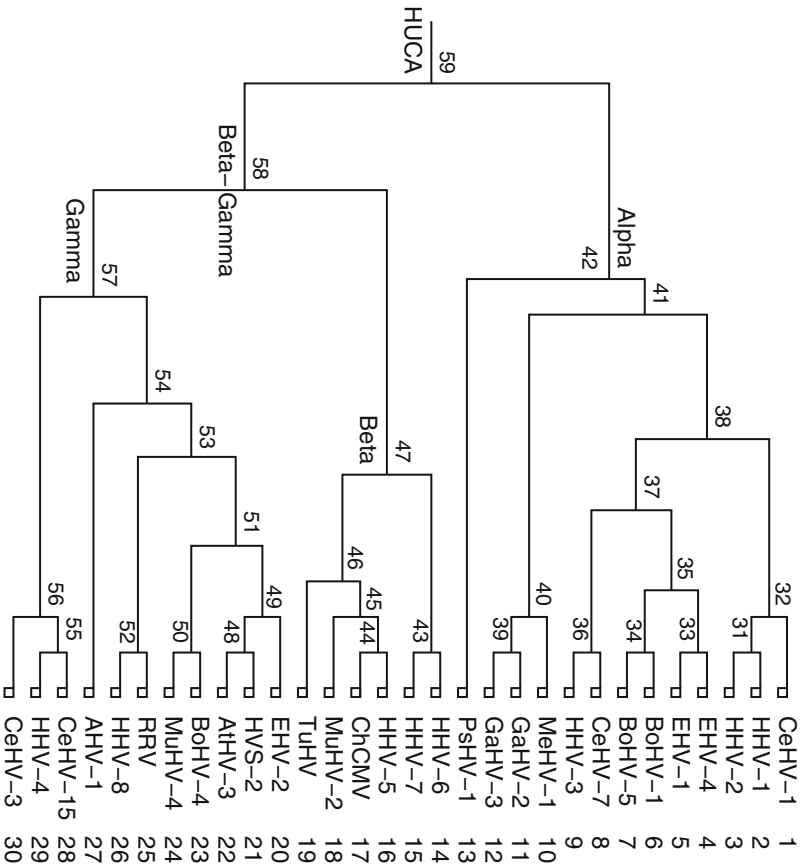


Fig. 1 Herpesvirus tree
Herpesvirus genomes evolutionary tree analyzed. The root corresponds to the herpesvirus ultimate common ancestor (HUCA); its child on the right to the ancestor of α superfamily, and the child on the left, to the common ancestor of β and γ superfamilies. The numbers are labels of different nodes on the tree

Unfortunately, the situation is less clear cut for the functionally similar proteins. Out of the 86 similar pairs available, there are 24 pairs (28%) that have their mutual similarity value less than 0.67. Thus at the similarity shift at 0.67, 28% of the similar pairs will not be identified as such, that is, at this similarity shift the method would lack sensitivity. To choose a similarity shift that minimizes the error in assigning negative and positive similarity values, one needs to compare the distribution of similarity values in the set of functionally similar pairs with that in the set of dissimilar pairs. As Fig. 2 shows, the graphs intersect when the similarity value *mbc* is 0.42.

Thus the external knowledge of functional similarity between some HPFs supplies us with two candidates for the similarity shift values, 0.67 and 0.42. There

Table 1 Some previously determined herpesvirus common ancestor D-HUCA's (Davison 2002; McGeoch et al. 2006) functions within membrane glycoproteins in the herpesvirus ancestor (two columns on the right) versus the results from the mapping of our clusters (three columns on the left); with function descriptions taken from VIDA

Mapping	H/APF	Description, gp – – glycoprotein	HSV-1 gene	D-HUCA
HUCA	20	gp M, HHV-1 UL10	UL10	gp M; compl. with gp N
HUCA	3	gp B, HHV-1 UL27	UL27	gp B
HUCA	APF 3:		UL22	gp H; compl. with gp L
	42	*gp H, HHV-1 UL22*		
	12	*gp H, HHV-8 ORF22*		
	531	*gp H, HHV-8 ORF22*		
ALPHA	47	gp L, HHV-1 UL1	UL1	gp L; compl. with gp H
BETA	50	gp L, HHV-5 UL115		
GAMMA	114	gp L, HHV-8 ORF47		
GAMMA	296	gp L, MuHV-4 ORF47		

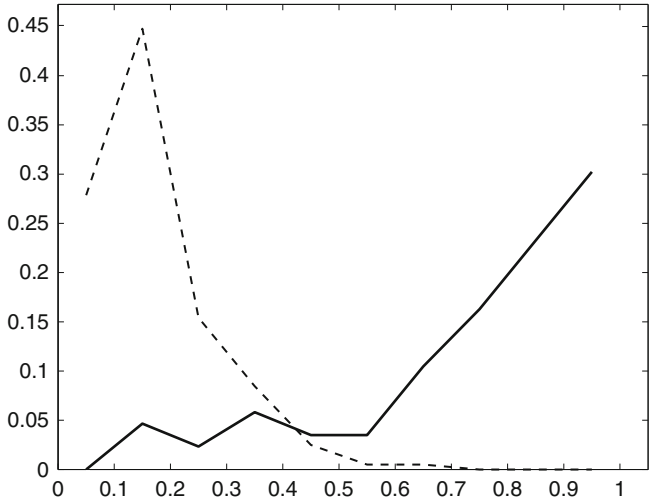


Fig. 2 Empirical frequency functions for the sets of functionally similar pairs (*solid line*) and dissimilar pairs (*dashed line*). The x-values represent the *mbc* similarity

are 80 APF 0.67-clusters comprising original 180 HPFs and leaving 560 HPFs unclustered, and 102 0.42-APF clusters over original 249 HPFs, and 491 HPFs unclustered. The first 80 0.42-clusters correspond one-to-one to the 80 0.67-clusters. Which one is more suitable? To answer this, we are going to develop and use more knowledge of the genomes.

4 Advancing Genome Knowledge

4.1 Reconstructed Histories of HPFs

For the further analysis, we utilize the evolutionary histories of HPFs over the evolutionary tree. These histories have been derived using the principle of maximum parsimony (Mirkin et al. 2003). These histories supply us with the reconstructed HPF contents of all the genome ancestors according to the tree. Of these, currently most useful are reconstructions of the most ancient genomes, those of ancestors of superfamilies α, β and γ, as well as the more universal common ancestors, HUCA and $\beta\gamma$. This is because the similarities and differences among herpesvirus species are somewhat better understood at deeper levels.

The reconstructions of the five ancestors with APFs found at the two similarity shift values, $\lambda_0 = 0.42$ and $\lambda_0 = 0.67$, are essentially the same. The only exception is the common ancestor of the α superfamily, which gains three more APFs when λ_0 decreases from 0.67 to 0.42. These are: (1) APF81 comprised of HPFs 9 and 504, both of glycoprotein C; (2) APF82 comprised of HPF 38 and HPF 736, both of glycoprotein I; and (3) APF84 comprised of HPF 47 and HPF 205, both of glycoprotein L. Unfortunately, at the current state of domain knowledge, we cannot interpret the phenomenon of simultaneously gaining the three glycoprotein families in terms of the α herpesvirus activities alone.

We can, however, look at the mutual positions of genes bearing these proteins within the virus genomic circular structures. We find that in all 13 genomes comprising α superfamily in our data, gene bearing glycoprotein E always immediately precedes that of glycoprotein I. This by itself may be considered a strong indication that there must be a mechanism in the superfamily involving both glycoproteins that has been developed already in the α ancestor. Moreover, it appears, glycoprotein E corresponds to an aggregate protein family comprised of HPF 26 and HPF 301 (at both levels of the similarity shift, 0.67 and 0.42) that has been mapped by our algorithm to the ancestral α node (Mirkin et al. 2006). This leads us to conclude that glycoprotein I must also belong to the α ancestor, thus implying that similarity shift $\lambda_0 = 0.42$, better fits to the knowledge added by the reconstruction than $\lambda_0 = 0.67$, because in which glycoprotein I's aggregate family falls in α ancestor only at the former value.

4.2 Derived Ancestors of Herpes Proteins

The analysis of glycoproteins in the α superfamily has led us to accept the value $\lambda_0 = 0.42$ and the corresponding number of protein families, after aggregation, 593. Some of the structural conclusions from the mapping of the aggregate 0.42-families to the evolutionary tree are presented in Table 1 taken from Mirkin et al. (2010).

The common ancestor of herpesviruses, HUCA, according to our reconstruction, should be comprised of 29 protein families. These all are well studied proteins except only three of the participating families of no known function.

Relations between our mapping results and D-HUCA are illustrated in Table 1: the fragmented HPFs, having been aggregated into APF3, fall into HUCA, yet some HPFs clearly fail to aggregate (47, 50, 114 and 296). The ancestor of each α-, β-, and γ family, has a glycoprotein L, so that the corresponding gene may have been present in HUCA as well. The HPFs have no significant sequence similarity nor common neighbors and, thus, cannot be combined together by clustering alone. Yet, at the genome organisation level each of the glycoprotein L genes always exactly precedes the corresponding Uracil-DNA glycosylase gene, which is mapped into HUCA, according to our reconstruction. This suggests that these are common ancestral genes indeed; just they have undergone sequence change to a level where sequence similarity is no longer sufficient to assign homology.

Concerning other four superfamily ancestors in our study, α, $\beta\gamma$, β and γ, our reconstructions show that only the contents of the α superfamily is relatively well studied. This means that the mechanisms separating the three superfamilies, especially those for β and γ, are yet to be investigated. Our reconstructed histories give clear indications of what proteins should be studied next.

5 Conclusion

Clustering is an activity purported to help in enhancing knowledge of the area the data relate to. Typically, this comes via a set of features assigned to the entities; the features reflect the knowledge and are to be used in interpreting cluster results. In proteomic studies, entities are frequently supplied with their similarities only, lacking any sensible features to look at when interpreting results. In such a situation, data recovery clustering supplies a reasonable device for reflection of the knowledge of proteome, the similarity shift value. Using two sets of protein pairs, those that should and those that should not fall into the same clusters, may lead to considerably narrowing down the choice of reasonable shift values, as shown above. One more step is in using the parsimonious reconstruction of the evolutionary history of the clusters. This may allow both further reduction of choices by confronting the reconstructions with the gene arrangement within genomes and interpretation of the clusters.

A possible direction for further work can be application of similar principles for clustering and interpreting of protein families at other sets of related genomes.

References

Alba, M. M., Lee, D., Pearl, F. M., Shepherd, A. J., Martin, N., Orengo, C., & Kellam, P. (2001). VIDA: A virus database system for the organisation of animal virus genome open reading frames. *Nucleic Acid Research, 29*, 133–136.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research, 25*, 3389–3402.

Bader, G. D., & Hogue, C. W. V. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics, 4*, 2.

Ben-Dor, A., Shamir, R., & Yakhini, Z. (1999). Clustering gene expression patterns. *Journal of Computational Biology, 6*, 281–297.

Brown, D. P., Krishnamurty, N., & Sjolander, K. (2007). Automated protein subfamily identification and classification. *PloS Computational Biology, 3*(8), e160, 1526–1538.

Chen, Y., Reilly, K. D., Sprague, A. P., & Guan, Z. (2006). SEQOPTICS: A protein sequence clustering system. BMC Bioinformatics, 7(Suppl. 4), S10.

Davison, A. J. (2002). Evolution of the herpesviruses. *Veterinary Microbiology, 86*, 69–88.

Felsenstein, J. (2001) *PHYLIP 3.6: Phylogeny Inference Package*. http://evolution.genetics.washington.edu/phylip/.

Gouzy, J., Eugene, P., Greene, E. A., Khan, D., & Corpet, F. (1997). XDOM, a graphical tool to analyse domain arrangements in any set of protein sequences. *Computer Applications in the Biosciences, 13*, 601–608.

Jarvis, R. A., & Patrick, E. A. (1973). Clustering using a similarity measure based on shared nearest neighbors. *IEEE Transactions on Computers, 22*, 1025–1034.

Kawaji, H., Takenaka, Y., & Matsuda, H. (2004). Graph-based clustering for finding distant relationships in a large set of protein sequences. *Bioinformatics, 20*(2), 243–252.

McGeoch, D. J., Rixon, F. J., & Davison, A. J. (2006). Topics in herpesvirus genomics and evolution. *Virus Research, 117*, 90–104.

Mirkin, B. (1976). *Analysis of categorical features*. Moscow: Statistika Publishers (in Russian).

Mirkin, B. (1987). Additive clustering and qualitative factor analysis methods for similarity matrices. *Journal of Classification, 4*, 7–31; Erratum (1989), 6, 271–272.

Mirkin, B. (1996). *Mathematical classification and clustering*. Dordrecht: Kluwer Academic Press.

Mirkin, B., Fenner, T., Galperin, M., & Koonin, E. (2003). Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evolutionary Biology, 3*, 2 (www.biomedcentral.com/1471-2148/3/2/).

Mirkin, B., Camargo, R., Fenner, T., Loizou, G., & Kellam, P. (2006). Aggregating homologous protein families in evolutionary reconstructions of herpesviruses. In D. Ashlock (Ed.), *Proceedings of the 2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology* (pp. 255–262). Piscataway, NJ.

Mirkin, B., Fenner T., Camargo R., Loizou G., & Kellam P. (2010) Similarity clustering of proteins using substantive knowledge and reconstruction of evolutionary gene histories in herpesvirus, Theoretical Chemistry Accounts: Theory, Computation, and Modeling, *125*, 3–6, 569–581.

Paccanaro, A., Casbon, J. A., & Saqi M. (2006). Spectral clustering of protein sequences. *Nucleic Acids Research, 34*, 1571–1580.

Poptsova, M. S., & Gogarten, J. P. (2007). BranchClust: A phylogenetic algorithm for selecting gene families. *BMC Bioinformatics, 8*, 120.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities by overlapping properties. *Psychological Review, 86*, 87–123.

Tatusov, R. L., Galperin, M. Y., Natale, D. A., & Koonin, E. V. (2000). The COG database: A tool for genome-scale analysis of protein function and evolution. *Nucleic Acids Research, 28*(1), 33–36.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research, 22*, 4673–4680.

Microarray Dimension Reduction Based on Maximizing Mantel Correlation Coefficients Using a Genetic Algorithm Search Strategy

Elena Deych, Robert Culverhouse, and William D. Shannon

Abstract We present the GA-Mantel algorithm to find in high dimensional microarray data a subset of genes that captures relevant spatial relationships among the samples, in order to reduce the data for further analysis and eventually to identify meaningful biological markers. GA-Mantel uses a genetic algorithm to search over possible probe subsets using the Mantel correlation as the scoring measure for assessing the quality of any given probe subset, and consensus methods for selecting the final list of important genes. GA-Mantel is evaluated on both artificial data sets and on experimental microarray data taken from leukemia patients. Current results indicate the GA-Mantel method exhibits promise as a way of efficiently identifying information-rich gene subsets in large data sets while avoiding the curse of dimensionality.

1 Introduction

A major stumbling block in analyzing high throughput array data is that there are too many genes (probes) for the number of samples. Datasets with a small number of samples (N) relative to a large number of measurements or variables (P) belong to the class of problems known as the large P small N problem, or more simply the 'curse of dimensionality' (Bellman 1961). The 'curse of dimensionality' means that fitting standard statistical models and making accurate predictions gets very hard as the number of variables (dimensions) increases. The implication for array data is that, due to the high number of variables, we can expect any models fit to these data to be inaccurate.

Three nearly equivalent statements about the curse of dimensionality are that (a) in the high dimensional space, samples are very far from one another so that accurate descriptions of the data distribution becomes impossible (sparseness);

W. D. Shannon (✉)
Washington University School of Medicine, St. Louis, MO, USA
e-mail: wshannon@wustl.edu

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_5,

(b) there are many more possible interaction terms, nonlinear effects, etc. to consider so that there quickly become too many possible models to test (model complexity); and (c) gene subsets are randomly correlated with phenotype leading to numerous spurious 'significant' results in the data that may mask the true correlations between genes and phenotype or other outcomes (random multicollinearity). Sparseness, model complexity, and multicollinearity can result in finding many genes that appear to be associated with phenotype but which in reality are associated by chance alone and not verifiable in follow-up studies. These genes can lead to statistical models with perfect or near perfect ability to classify patients into phenotype subgroups even when there is no 'true' relationship between the genes and the phenotype.

While nonparametric multivariate regression methods have been proposed for finding significant genes in microarray data (e.g., CART, neural nets), these approaches require a lot of model searching and therefore use up degrees of freedom rapidly (a problem for the comparatively small sample sizes). As a result, there is little or no information left to determine if the chosen model is statistically significant. Cluster analysis uncovers structure in data and is a statistical tool commonly used in microarray data analysis. However, for the *large P small N problem* there are often many distinct sets of clusters, arising completely by chance, that can perform optimally for almost any measure of goodness-of-fit. In such situations, it is impossible to decide the best cluster model. Classification models predict a sample's group membership (i.e., is the sample tumor or normal tissue) from the variables. However, these methods require training datasets to fit the model and a validation dataset to assess its classification accuracy, thus accentuating the problems related to small sample sizes.

Likewise, resampling (cross validation, bootstrapping), model averaging (bagging), or iterative reweighting (boosting) (Kerr and Churchill 2001; Breiman 2005; Dettling 2004; Jiang et al. 2008) is usually ineffective for large *P small N data*. For example, multicollinearity in the full dataset will be present in the bootstrapped samples as well so model fitting and model accuracy estimates still are subject to the problems of the curse of dimensionality.

Several gene selection strategies are used to reduce the number of genes that need to be analyzed. The simplest approach uses gene filtering based on thresholds (Pounds and Cheng 2005). While useful, threshold methods are ad hoc which make them less attractive to statisticians. Recent work on dimension reduction using classical multivariate methods such as PCA and PLS has been developed and tested (Dai et al. 2006). In these approaches a small number of linear combinations or projections of the original genes are used to replace the gene list such that most of the variability in the data is retained. Classifying samples based on this subset of linear combinations performs well, though the effects of individual probes on classification have to be discerned from their coefficient weights in the linear combinations.

Overall, there is currently a great need for developing new methods of analyzing array data.

2 Methods

Mantel Correlation We first described the use of Mantel statistics in microarray data analysis (Shannon et al. 2002). Mantel statistics were developed in 1967 to correlate temporal and spatial distributions of cancer incidences (Mantel 1967) and extended in 1987 to the partial correlation and regression framework (Smouse and Long 1986). The basic idea is to transform standard data matrices (i.e., subject by covariate data) into subject pairwise distances or similarities and to analyze these proximity matrices instead of the raw data. The result is that instead of analyzing an $N \times P$ data matrix where the *large P small N problem* exists, we analyze a smaller $N \times N$ matrix format. In 2002 we used Mantel statistics to correlate expression profiles of multiple genes simultaneously, as opposed to one gene at a time, with patient brain tumor phenotype (Watson et al. 2001). We also ranked gene subsets by their relationship to subject phenotype.

Mantel correlation is calculated as follows. Consider a microarray dataset on N subjects with P genes. Calculate an $N \times N$ subject pairwise proximity (distance or similarity) matrix by any appropriate metric such as the Euclidean distance (Eisen et al. 1998; Shannon et al. 2003). Let $S_i, i = 1, \ldots, N$ denote the subjects, G_j, $j = 1, \ldots, P$ denote the genes, and $x_{i,j}$ denote the expression of gene G_j in subject S_i, then the Euclidean distance between two subjects i, i' is (1)

$$d_{i,i'} = \sqrt{(x_{i,1} - x_{i',1})^2 + (x_{i,2} - x_{i',2})^2 + \cdots + (x_{i,P} - x_{i',P})^2}. \tag{1}$$

The more similar the gene expression profiles in two subjects, the smaller their pairwise distance (larger similarity), and vice versa. Thus, the $N \times N$ raw data is transformed into the subject pairwise distances matrix D transforming array data from a *large P small N problem* to a more tractable data analysis problem (2):

$$\begin{array}{c} \textit{Sample} \\ 1 \\ 2 \\ 3 \\ \vdots \\ N \end{array} \begin{array}{c} \begin{array}{cccc} G_1 & G_2 & \ldots & G_P \end{array} \\ \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,P} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,P} \\ x_{3,1} & x_{3,2} & \ldots & x_{3,P} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \ldots & x_{N,P} \end{bmatrix} \end{array} \Rightarrow \quad D = \begin{bmatrix} 0 & d_{1,2} & d_{1,3} & \ldots & d_{1,N} \\ & 0 & d_{2,3} & \ldots & d_{2,N} \\ & & 0 & \ldots & d_{3,N} \\ & & & \ddots & \vdots \\ & & & & 0 \end{bmatrix} \tag{2}$$

where $d_{1,2}$ is the distance between samples 1 and 2, $d_{1,3}$ is the distance between samples 1 and 3, etc.

Similarly, for a phenotype dataset, a subject pairwise distance matrix can be calculated using an appropriate distance metric. We denote the phenotype distance matrix as D^{Pheno}(3)

$$\begin{array}{cc} \textit{Sample} & \textit{Phenotype} \\ 1 & \\ 2 & \\ 3 & \\ \vdots & \\ N & \end{array} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{bmatrix} \Rightarrow \quad D^{Pheno} = \begin{bmatrix} 0 & d_{1,2}^{Pheno} & d_{1,3}^{Pheno} & \dots & d_{1,N}^{Pheno} \\ & 0 & d_{2,3}^{Pheno} & \dots & d_{2,N}^{Pheno} \\ & & 0 & \dots & d_{3,N}^{Pheno} \\ & & & \ddots & \vdots \\ & & & & 0 \end{bmatrix} \tag{3}$$

Mantel correlation is calculated as the Pearson correlation coefficient between the two vectors of pairwise distances $(d_{1,2}, d_{1,3}, \dots, d_{N-1,N})$ and $(d_{1,2}^{Pheno}, d_{1,3}^{Pheno}, \dots, d_{N-1,N}^{Pheno})$ with statistical significance calculated by permutation testing to overcome the lack of independence between the pairwise distances within a matrix. Since the pairwise distance vectors are aligned to match the same subject pairs, the Mantel correlation indicates if the genes $G_1, G_2, \dots, G_P$ results in the same spatial distribution or pattern of separation among the samples (same relative distance matrix) defined by the phenotype. **A positive Mantel correlation indicates the genes in** (2) **contain the information for separating, discriminating, or predicting samples type (phenotype) in** (3). It is this property of the Mantel correlation that we propose to use to improve gene subset selection in array data.

For any given subset of genes, a distance matrix can be calculated as above. Let Q indicate the set of genes we are interested in, and assume in general that the number of genes, m, in Q will be significantly smaller than P. In practice, m might be on the order of a few dozen or few hundred genes that could be studied further in the lab. We can calculate a pairwise distance matrix on the same N samples using only this subset of genes $G_{(1)}, G_{(2)}, \dots, G_{(m)}$, $m \ll P$, where the subscript parentheses indicate these may be different from genes $G_1, G_2, \dots, G_m$ used in calculating the distance matrix D, to obtain a second distance matrix D^Q where

$$d_{i,i'}^{Q} = \sqrt{\left(x_{i,(1)} - x_{i',(1)}\right)^2 + \left(x_{i,(2)} - x_{i',(2)}\right)^2 + \dots + \left(x_{i,(m)} - x_{i',(m)}\right)^2}$$

Conceptually we think of the P genes in the entire gene list as consisting of signal and noise genes. The signal genes are responsible for how the samples separate into phenotype subgroups while the noise genes have little effect on the separation of the samples and would have no correlation with phenotype – the noise genes add error to the true pairwise distances $d_{i,i'}$, which in the perfect world the error would be distributed normally with mean 0, though we have not yet investigated this. If we can calculate D^Q using only signal genes the distances should be highly correlated to the distances in D^{Pheno}, and the Mantel correlation will be high. Distances calculated on a subset of noise genes, on the other hand, should have near-zero correlation with D^{Pheno}.

Mantel Correlation without Phenotype We described how Mantel correlation could be used to score gene subsets in array data that may be related to phenotype. Another important role of Mantel correlation is to find gene subsets that correlate with overall

patterns of gene expression. In this problem we calculate the distance matrix D on the full set of genes and D^{Q} on a subset of genes and correlate D with D^{Q}. The goal of this analysis is to separate signal from noise genes.

Genetic Algorithms Genetic algorithms are powerful search approaches for computational problems where approximate or exact solutions are hard or impossible to calculate (Grefenstette et al. 2005; Forrest 1993). Genetic algorithms apply computational 'evolutionary operators' to a population of possible solutions (i.e., population of gene subsets) to create the next generation of solutions (offspring) based on a measure of fitness which is evaluated on each solution. In this paper we propose using Mantel correlation as this measure of fitness, thus calling the method GA-Mantel. Genetic algorithms have been implemented in several standard statistical packages including SAS's experimental Genetic Algorithm module in SAS/IML, which we used to develop the GA-Mantel program.

In the beginning of the GA process, an appropriate representation of the search space must be specified. For gene selection from microarray data, solutions are lists or subsets of genes, and two members of the solution space might be:
where the first list specifies that genes 10, 123, 456, and so on, are in the gene subset, and the second list specifies that genes 29, 378, 456, and so on, are in the gene subset. Each solution is then used to generate a distance matrix D^{Q} on $G_{(1)}, G_{(2)}, \ldots, G_{(m)}$ where for the first solution $G_{(1)} = G_{10}$, $G_{(2)} = G_{123}, \ldots,$ $G_{(7)} = G_{923}$. The Mantel correlation is calculated using the distance matrix from the sample phenotypes D^{Pheno} or from other genes D. In order to introduce new candidate gene list solutions, recombination and mutation evolutionary operators are applied to parent solutions using the rank-based selection algorithm: recombination involves swapping features between solutions, and mutation involves a random change in a solution:

Gene subsets selected as parents for generating offspring solution is based on current goodness-of-fit. Those solutions with highest fitness (i.e., highest Mantel correlation) are preferentially selected in the hope that the genes in these solutions that result in higher Mantel correlation will get matched with other signal genes from mutation and crossover resulting in more improvement in Mantel correlation.

In our GA-Mantel program some parameters, such as the mutation and crossover rates, use SAS default values. However, three main parameters impacting how much of the solution space is searched must be defined: number of generations, solution length, and population size. The generation specifies how many times the GA iterates (i.e., can evolve) to find an optimal solution. The solution length specifies the number of genes to retain in calculating D^{Q}. The larger the solution length the more likely signal genes will be found. The population size specifies the number of solutions in a generation that will be evaluated. Increasing this parameter increases the coverage over the solution space and the variability introduced by evolutionary operators. Each of these parameters add to computing time as they increase, so in applications there will be a need to balance the search algorithm speed with the amount of the solution space searched.

GA-Mantel SAS Program GA-Mantel is fully implemented in SAS using SAS's experimental Genetic Algorithm module in SAS/IML, and is available for free from the authors.

Consensus method for identifying optimal gene subset We have seen in our applications that any given GA-mantel solution has genes that individually have high Mantel correlation and genes that individually have low correlation. Additionally, we have found that running GA-Mantel multiple times on the same data set results in solutions where some genes appear in many of the runs and most genes appear in few of the runs. The algorithm initialization and selection of the starting solutions (i.e., generation 1) is always random so no two solutions are likely to be to have 100% agreement. Since genes with high Mantel correlations are more likely to be repeatedly selected by GA-Mantel in independent runs, selecting genes appearing most often in solutions should enrich for signal genes of interest. We use a simple consensus method of keeping those genes that appear in more than some pre-specified threshold minimum number of runs, or those that have frequency of appearance that clearly separates them from the rest (i.e., 10 genes appear in $>30\%$ solutions, while all other appear in $<10\%$ solutions).

3 Results

We report on four analyses using GA-Mantel in this section. First, are the results of a simulation study to assess how changing parameters in GA-Mantel impacts the selection of signal genes in simulated data. Second, we applied GA-Mantel to the publically available Golub leukemia microarray data (Golub et al. 1999). Third, we applied GA-Mantel to two leukemia datasets collected at Washington University.

Evaluation on simulated data: Initial evaluation of the GA-Mantel method was performed on simulated data and reported previously (Grefenstette et al. 2005). In that report we showed that as GA-Mantel iterated through generations, the Mantel correlation increased, and that by increasing the search time (e.g., increasing the number of generations) GA-Mantel found more signal genes. In this paper we report on a more comprehensive study on simulated data, each with 1,000 variables (genes) and 40 samples (microarrays). For the genes, 10 (1%) were designated as *signal* genes, and the other 990 were *noise* genes. The experiments were derived from two groups of 20 samples each. In Group 1 (control), the 10 signal genes were sampled independently from a normal N (0,1) distribution. In Group 2 (cases), the signal genes were samples from one of a N (0.5, 1), N (1, 1), N (2, 1) or N (5, 1) distribution. All 990 noise genes in both groups were sampled from a normal distribution N (0, 1). GA-Mantel method was applied to these data sets using combinations of the following parameters: population size 100, 250, and 500; generations (iterations) 100, 250, and 500; solution length 10, 30, and 100. We simulated 100 independent datasets for each of the combinations of simulation scenario and GA-Mantel parameters.

Table 1 Mean number of signal genes out of 10 found

Mean Diff.	Iters	Population size								
		100			250			500		
		Solution length			Solution length			Solution length		
		10	30	100	10	30	100	10	30	100
0.5	100	0.9	1.5	3.3	0.9	1.9	3.3	1.0	2.0	3.6
	250	0.7	1.8	3.4	0.9	1.9	3.5	1.3	1.9	3.6
	500	0.8	1.9	3.4	0.9	2.0	3.7	1.0	2.0	3.9
1	100	3.0	5.0	7.2	4.1	6.2	8.3	4.3	6.7	8.7
	250	3.0	5.5	7.8	4.1	6.2	8.2	4.6	6.1	8.4
	500	3.2	5.5	8.0	4.0	6.0	8.3	4.3	6.7	8.3
2	100	4.6	7.6	9.7	6.4	9.1	10	7.7	9.5	10
	250	4.9	8.0	9.8	6.6	9.2	10	7.6	9.4	10
	500	5.1	8.6	9.9	6.9	9.4	10	7.6	9.3	10
5	100	4.7	7.8	9.8	6.8	9.0	9.9	8.1	9.3	9.9
	250	4.9	8.2	9.8	6.8	9.2	9.9	8.0	9.2	9.9
	500	5.2	8.5	9.8	7.2	9.2	9.8	8.2	9.2	9.8

For each simulation we collected and averaged across the 100 independent datasets to calculate two algorithm performance measures: the true positive values (number of true signal genes selected in final solution, out of 10) and Mantel correlation. As the GA parameters (number of generations, solution length, and number of iterations) increase, the number of signal genes captured increases (TP value gets higher), Table 1. In addition, the mean Mantel correlation increased from 0.40 to 0.99 for the final solutions (data not shown),

Since individual GA-Mantel solutions did not find high percentage of signal genes when the difference between cases and controls in signal genes was small and since all solutions included false positive (noise) genes, we attempted to improve the selection and reduce the amount of noise by the consensus process. We applied the consensus method by running GA-Mantel algorithm 100 times on a simulated dataset generated as described above with mean difference of 0.5 between the case and control groups in 10 signal genes. We counted the number of times each gene was found in 100 solutions. Consensus analysis resulted in clear-cut separation of noise and signal genes: Each signal gene appeared on average 37.9 out of 100 times and never appeared less than 33 times out of 100. No noise gene appeared more than 21 times out of 100. In this analysis the signal genes could be distinguished from noise genes by setting up a cut off point for consensus of appearing in at least 33 out of 100 independent runs. Such cut off point, although ad hoc is justifiable given the large gap in prevalence of signal and noise genes. Presumably with more runs of GA-Mantel and improved consensus methods enrichment for signal genes can be improved.

We then compared the results of the GA-Mantel consensus analysis with a standard parametric t-test analysis. We generated 1,000 independent artificial datasets using the criteria in the previous paragraph and captured the p-values for all signal and noise genes in each dataset. For this analysis, we defined true positive (TP) as

the number of signal genes and False positive (FP) as the number of noise genes that were significantly different between cases and controls in a t-test. In 1,000 datasets, the mean TP values were 3.3, 1.4, and 0.3 and mean False Positives were 48.9, 9.4, and 0.9 at p-values of 0.05, 0.01, and 0.001 respectively. This result indicates that using a liberal p-value of 0.05 results in finding on average only 3 out of 10 signal genes and incorrectly identifying 49 noise genes as significant (signal). Using a more conservative p-value of 0.001 (i.e., adjusting for multiple testing) results in elimination of most false positives but at a cost of losing almost all signal genes (TP = 0.3). Therefore, in this simulation, when the signal is low, the t-test cannot identify signal genes or eliminate a large number of noise genes successfully, regardless of p-value specifications. On the other hand, the GA-Mantel in conjunction with consensus method was able to successfully separate noise genes from signal genes.

Golub Leukemia Data: To illustrate GA-Mantel correlation with real data we applied it to the publicly available Golub microarray leukemia data set consisting of 7,129 genes measured on 27 acute lymphoblastic leukemia or ALL (samples 1–27) and 11 acute myeloid leukemia (samples 28–38) patients (Golub et al. 1999). In Fig. 1, the left dendrogram used a distance matrix measured on all 7,129 genes and produces a good separation of the two leukemia types with 18 ALL and 1 AML in the left cluster (branch) and 9 ALL and 10 AML in the right cluster. The right dendrogram used a distance matrix calculated on 50 genes found by the GA-Mantel algorithm and produces nearly the same dendrogram as that produced with all 7,129 genes. This indicates that this subset of 50 genes contain much of the same spatial relation information as is found in all 7,129 genes. It was also found (data not shown) that fitting dendrograms to 50 random genes resulted in completely different trees with no ability to separate the two phenotype subgroups and near 0 Mantel correlation.

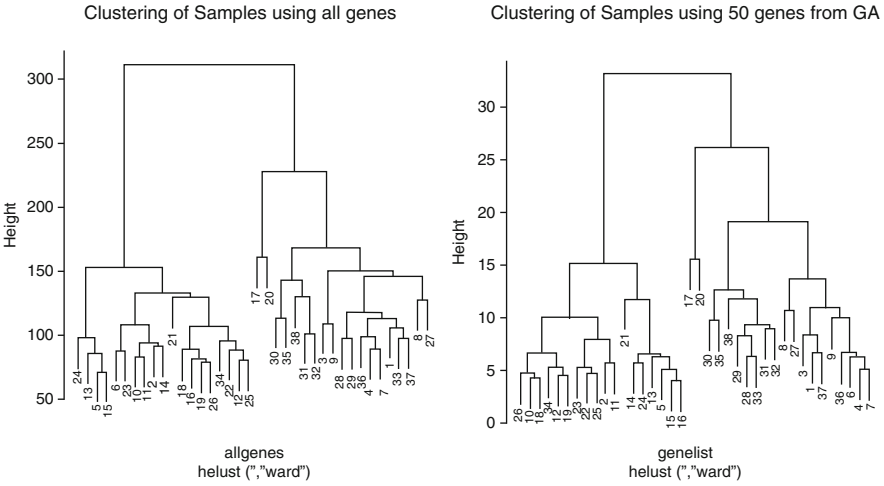


Fig. 1 Golub leukemia data

Analysis of Chemotherapy Resistance Microarray Data: Mobilization of HSCs (hematopoietic stem cells) for treatment of leukemia in mice and humans can be induced using the drug G-CSF (G) or by the use of chemokines and chemokine receptor antagonists. Recent preclinical and clinical data using the drug bicyclam AMD3100 (A) suggest that the combination of G+A results in significantly improved yields of HSCs compared to G alone in both mice and humans (Flomenberg et al. 2005). To identify genes that are differentially expressed following G or A mobilization, we performed RNA profiling analyses using Affymetrix U133+2 arrays and RNA isolated from purified (>95%) CD34+ HSCs on paired data obtained from eight individual normal donors mobilized with G followed later by a second mobilization with A. To reduce the number of genes to be analyzed we imposed two threshold rules. The first threshold is that a gene was retained if and only if it exceeded an expression of at least 500 in all the microarrays. This level of expression is believed to be well above the level of the microarray noise. The second threshold imposed was that the ratio of AMD-to-GCSF within each pair for the gene had to be at least a two-fold difference all in the same direction (i.e., the AMD-to-GCSF ratio >2 in all pairs or was <0.5 in all pairs). This reduced the number of genes for analysis to 18 that clearly showed a strong and consistent change in expression between the A and G treatment groups. This dual-threshold analysis was followed by correlating these 18 genes with the remaining genes to find other genes that did not meet the two-fold threshold requirements but had similar expression patterns in separating A and G samples as the 18 selected genes. The search, performed by GA-Mantel, identified 148 additional probes that produced a distance matrix correlated with the distance matrix produced by the 18 genes found by the threshold method. In Fig. 2, the dendrogram on the left is based on the 18 genes found by the threshold method and the dendrogram on the right is based on the 148 additional genes found by GA- Mantel. The samples are labeled as A1–A8 for AMD3100 and G1–G8 for G-CSF. Both dendrograms have two distinct branches that perfectly separate the samples from the two treatments. The dendrogram clearly indicates that GA-Mantel method resulted in identification of 148 additional probes that separate the treatment samples but could not be identified by the threshold method. (No adjustment for paired data has been developed yet.)

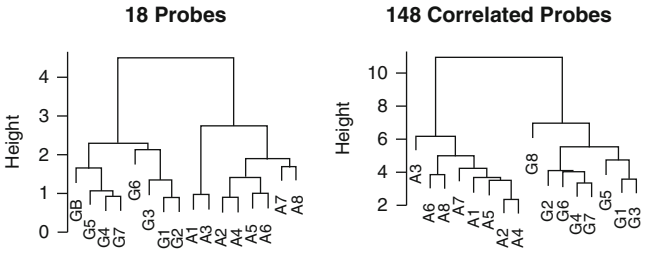


Fig. 2 Chemotherapy resistant data

Drug resistant APL cell lines: An experiment was done to try to reproduce cell drug resistance in a mouse leukemia cell line. Cell lines were created using banked leukemia cells isolated from mice that had a particular genetic abnormality. These mice develop a leukemia-like disease followed by death anywhere from 6 months to 2 years of age. Approximately 100 tumors have been banked to date with each representing an individual mouse from the colony. Fifteen individual cell lines have been established from these tumors, with six showing drug resistance to ATRA (retinoic acid), AR (arsenic trioxide) AraC (cytarabine) or DNR (daunorubicin / doxorubicin). Cell lines were microarrayed after being grown in ATRA (N = 3), AR (N = 2), AraC (N = 2), DNR (N = 2), and no drug (N = 1). Our goal was to identify genes that are expressed differently in the presence of different drugs in hopes that these may indicate biological processes conferring resistance.

From the design of the study the phenotype distance matrix D^{Pheno} was calculated based on whether the arrays were grown in the same drug (distance = 0) or different drugs (distance = 1). A GA-Mantel search for 500 probes resulted in a solution with Mantel correlation = 0.79 and a close reproduction of the phenotype reference distance matrix. It is interesting to note that the dendrogram fit to the all gene expression data did not reproduce the experimental design (not shown). We then narrowed down the list using consensus where we ran GA-Mantel 100 times with solution length of 500 genes and selected only genes that appeared in all solutions. This resulted in seven genes that result in a Mantel correlation of 0.72 and produce a dendrograms showing good separation of the drug treatment groups (Fig. 3). In this dendrogram, the sample not treated by a drug is indicated as "ref", the sample treated by DNR are marked as "dn1" and "dn2", AR samples are marked "ar1" and "ar2", ATRA samples are "AT1, AT2, and AT3", and AraC are marked as "AA1" and "AA2".

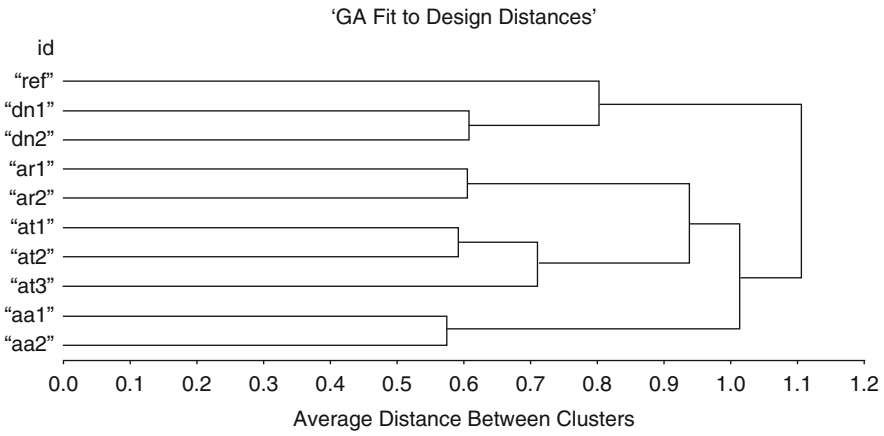


Fig. 3 Drug resistant cell lines

4 Discussion

GA-Mantel is a general purpose variable selection method for any cluster analysis problem. For our specific application in genetics, current results indicate that the GA-Mantel is a promising novel method of efficiently identifying information-rich gene subsets in large data sets. The analysis of artificial datasets indicates that by increasing algorithm parameters and performing consensus analysis, the signal genes can be found with the GA-Mantel method even when the difference between cases and controls is small and the parametric methods, such as t-test fail. Using a consensus method eliminates a large proportion of noise genes without decreasing the power to detect signal genes. In our artificial dataset analysis, there was a clear-cut separation in frequency of appearance between signal and noise genes: No signal genes appeared in less than 33 solutions and no noise gene appeared in more than 21 solutions. This means that any cut off point between 21 and 33 would result in perfect separation of signal from noise genes for this simulation study.

Of special note is the use of GA-Mantel for data with limited or no phenotype information. In our chemotherapy resistance data, the method identified 148 genes without considering phenotype but by correlating the total gene space with 18 genes found by a different method. These 148 genes were then found to clearly separate samples by phenotype.

Future direction: At the present time, the method is still a work in progress. When the true signal is small, the method requires large parameters for detection which can be computer-intensive. Several GA parameters used by SAS/IML Genetic algorithm are currently left as default and may require further investigation for program optimization. Selection of solution length is somewhat arbitrary at this stage and is determined mostly by practical (laboratory resources for re-testing likely signal genes) and not statistical reasons. We are also working on implementing the algorithm for more sophisticated genetic problems, such as epistasis, genomewide association data, and metagenomic data.

Acknowledgements We thank Dr. John DiPersio, Dr. Mike Rettig, and Mr. Matthew Holt for access to their leukemia microarray data and for their close collaboration over the years. This work was partially supported by an NIH Program Project Grant (5P01CA101937, Genomics of Acute Myelogenous Leukemia) and the Washington University Dept. of Medicine's Biostatistical Consulting Center Directed by Dr. Shannon.

References

Bellman, R. E. (1961). *Adaptive Control Processes*. Princeton, NJ: Princeton University Press.

Breiman, L. (2005). *Bagging Predictors Machine Learning, 24*, 123–140.

Dai, J. J., Lieu, L., & Rocke, D. (2006). Dimension reduction for classification with gene expression microarray data. *Statistical Applications in Genetics and Molecular Biology, 5*.

Dettling, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics, 20*, 3583–3593.

Eisen, M. B., et al. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences, 95* (pp. 14863–14868).

Flomenberg, N., et al. (2005). The use of AMD3100 plus G-CSF for autologous hematopoietic progenitor cell mobilization is superior to G-CSF alone. *Blood, 106*, 1867–1874.

Forrest, S. (1993). Genetic algorithms: Principles of natural selection applied to computation. *Science, 261*(5123), 872–878.

Grefenstette, J., Thompson, K., Shannon, W., & Steinmeyer, B. (2005). Genetic algorithms for feature selection using Mantel correlation scoring. *Interface 2005: Classification and Clustering, 37th Symposium on the Interface*. St. Louis, MO.

Golub, T. R., et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science, 286*(5439), 531–537.

Jiang, W., Varma, S., & Simon, R. (2008). Calculating confidence intervals for prediction error in microarray classification using resampling. *Statistical Applications in Genetics and Molecular Biology, 7*.

Kerr, M. K., & Churchill, G. A. (2001). Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences, 98*, 8961–8965.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research, 27*, 209–220.

Pounds, S., & Cheng, C. (2005). Statistical development and evaluation of microarray gene expression data filters. *Journal of Computational Biology, 12*, 482–495.

Shannon, W. D., et al. (2002). Mantel statistics to correlate gene expression levels from microarrays with clinical covariates. *Genetic Epidemiology, 23*, 87–96.

Shannon, W., Culverhouse, R., & Duncan, J. (2003). Analyzing microarray data using cluster analysis. *Pharmacogenomics, 4*, 41–52.

Smouse, P., & Long, J. (1986). Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic Zoology, 35*, 627–632.

Watson, M. A., et al. (2001). Gene expression profiling with oligonucleotide microarrays distinguishes World Health Organization grade of oligodendrogliomas. *Cancer Research, 61*, 1825–1829.

Part II
Classification and Data Analysis

Multiparameter Hierarchical Clustering Methods

Gunnar Carlsson and Facundo Mémoli

Abstract We propose an extension of hierarchical clustering methods, called *multiparameter hierarchical clustering methods* which are designed to exhibit sensitivity to density while retaining desirable theoretical properties. The input of the method we propose is a triple (X, d, f), where (X, d) is a finite metric space and $f : X \to \mathbb{R}$ is a function defined on the data X, which could be a density estimate or could represent some other type of information. The output of our method is more general than dendrograms in that we track two parameters: the usual scale parameter and a parameter related to the function f. Our construction is motivated by the methods of *persistent topology* (Edelsbrunner et al. 2000), the Reeb graph and Cluster Trees (Stuetzle 2003). We present both a characterization, and a stability theorem.

1 Introduction

Clustering techniques play a very central role in various parts of data analysis. They can give important clues to the structure of datasets, and therefore suggest results and hypotheses in the underlying science. However, despite being one of the most commonly used tools for unsupervised exploratory data analysis, and despite its extensive literature, very little is known about the theoretical foundations of clustering methods. These points have been made prominent by Ben-David and von Luxburg in Ben-David et al. (2006).

The general question of which methods are "best", or most appropriate for a particular problem, or how significant a particular clustering is has not been addressed very frequently. In the context of standard clustering (standard clustering refers to clustering methods that output a single partition of a dataset and hierarchical methods that yield a nested family of partitions), J. Kleinberg proves in Kleinberg (2002)

F. Mémoli (✉)
Mathematics Department, Stanford University, Stanford, CA
e-mail: memoli@math.stanford.edu

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_6,

a very interesting impossibility result for the problem of even defining a clustering scheme with some rather mild invariance properties.

Inspired by Kleinberg's axiomatic treatment, in Carlsson and Mémoli (2008) we wondered whether in the context of hierarchical clustering (HC from now on) methods, one would be able to lift the obstruction to existence in his result. Interestingly, we were able to prove that for HC methods, conditions similar to Kleinberg's yield uniqueness instead of non-existence. This HC scheme singled out by our theorem satisfies precise *stability* and *convergence* properties (Carlsson and Mémoli 2008). This unique scheme turned out to be *single linkage* HC. There seems to exist an agreement that amongst hierarchical methods, SL is the one with best theoretical properties, see also the results of Jardine and Sibson in this respect (Jardine and Sibson 1971).

However, single linkage has frequently been severely criticized for the *chaining effect* it exhibits (see Lance and Williams 1967; Wishart 1969, p. 296): SL will disregard the density of samples in a region and may tend to connect two dense clusters when just a few isolated samples produce a chain connecting them. This has had the effect that in practice other clustering methods are typically preferred over SL. Practicioners tend to favour average (AL) or complete (SL) linkage, which are deemed more sensitive to variations of density in datasets. However, since AL and CL are actually *unstable* (Jardine and Sibson 1971, Sect. 7.4) in a precise sense, there is a blatant inconsistency between the conclusions of theoretical studies and practical applications of clustering algorithms.

Clustering can be regarded as a statistical problem if we consider the dataset $\mathbb{X} = \{x_1, \ldots, x_n\} \subset X$ as a sample from some unknown probability measure μ_X defined on the Borel sets of a metric space (X, d_X). Consider for the sake of simplicity that X is Euclidean space $\mathbb{R}^d$ and that μ_X is a measure with density ρ with respect to the n-dimensional Lebesgue measure. The two main statistical approaches to clustering are the *parametric approach* and the *nonparametric approach*. The former approach is based on the assumption that each group i is represented by a density ρ_i that is a member of some parametric family. The density ρ is then a mixture of the group densities, and the number of components in the mixture together with the parameters values are estimated from the data. The latter approach assumes that groups correspond to *modes* of the density ρ. Searching for modes as a manifestation of the presence of groups can be traced back to D. Wishart's paper Wishart (1969).

With regards to the chaining effect: it is well understood that one of the shortcomings of SL is its insensitivity to *density*. In this direction, a classical result of Hartigan (1981) proves that SL is not *consistent* in the sense that it is unable to recover modes of an underlying density in $\mathbb{R}^d$ for all d. In Wishart (1969) Wishart proposes *one level mode analysis* as an obvious approach to the amelioration of the chaining effect. The idea is to remove from the observational data all the points that appear to be noise. Define the superlevel set $L_\rho(\sigma)$ of a density ρ at level σ as the subset of the underlying space X for which the density exceeds σ: $L_\rho(\sigma) = \{x | \rho(x) > \sigma\}$. Then, if $\widehat{\rho}$ is some estimate of ρ and σ a given threshold, the idea consists of applying SL clustering to $L_{\widehat{\rho}}(\sigma)$.

In Hartigan (1975) [Sect. 11] and Hartigan (1981), Hartigan expanded on Wishart's idea and made it more precise: he defined the *high density clusters* at level σ as the connected components of $L_\rho(\sigma)$. Hartigan also pointed out that the collection of high density clusters has a *hierarchical structure*: for any two clusters A and B (possibly at different levels) either $A \subset B$ or $B \subset A$ or $A \cap B = \emptyset$. This hierarchical structure is summarized by the **cluster tree** of ρ.

More recent instantiations of the one level mode analysis idea can be found in Ester et al. (1996); Cuevas et al. (2001); Biau et al. (2007). Typically, methods roughly consist of four steps: (1) for each data point calculate a density estimate $\widehat{\rho}$; (2) choose a density threshold σ and construct $L_{\widehat{\rho}}(\sigma)$; (3) construct a graph interconnecting all observations in $L_{\widehat{\rho}}(\sigma)$ within distance ε of each other; (4) define the clusters to be the connected components of this graph.

As was pointed out in Stuetzle and Nugent (2008), a well known weakness of the one level mode analysis is that the degree of separation between connected components of $L_\rho(\sigma)$, and therefore of $L_{\widehat{\rho}}(\sigma)$, depends critically on the choice of the density threshold σ, which is left to the user. Moreover, there might not be a single value of σ that uncovers all the modes. In Wishart (1969), citing this difficulty, Wishart proposed *hierarchical mode analysis*, which can be regarded as a procedure for computing the cluster tree of a density estimate $\widehat{\rho}$. The work of Wong and Lane (1983) provides a method of estimating the cluster tree of a density by a construction based on k-nearest neighbor density estimates.

In Stuetzle (2003) Stuetzle gives a precise recursive definition of the cluster tree. Stuetzle's method estimates the cluster tree of the density by computing the cluster tree of the nearest neighbor density estimate and then pruning branches believed to correspond to spurious modes. In Stuetzle and Nugent (2008) the authors present a generalization of Stuetzle's method to other density estimates. It is already expressed in the work of Stuetzle and Nugent that it is desirable to prove that the cluster tree estimates one constructs are *stable* to perturbations in the data. Furthermore, the issue of convergence of the sample based cluster tree has to be resolved, see the discussion in Wong and Lane (1983). Similar ideas are also present in the work of Klemelä (2004).

The construction implicit in many of the methods we mentioned can be paraphrased as follows. Assume (X, d_X, f) is given where (X, d_X) is a finite metric space and $f : X \to \mathbb{R}$ is a given function (which could be a density estimate). For each σ let $X^\sigma := L_f(\sigma)$. For a given $\varepsilon > 0$ consider the graph $G_{\varepsilon,\sigma} = (X^\sigma, E_{\varepsilon,\sigma})$ with $E_{\varepsilon,\sigma} = \{(x, x') \in X^\sigma \times X^\sigma \,|\, d_X(x, x') \leq \varepsilon,\ i \neq j\}$. Then, obtain a one-mode-analysis type of summary by computing the connected components of $G_{\varepsilon,\sigma}$. Clearly, this set up can be used for estimating the cluster tree of f as well by following a recursive procedure such as the one delineated by Stuetzle.

The proposal in this paper hinges on the idea that there is more information contained in the whole collection of graphs $\{G_{\varepsilon,\sigma}\}_{\varepsilon\geq 0,\sigma\geq 0}$ than just an estimate or a family of estimates (one for each ε) of the cluster tree. Much in the same way as single mode analysis suffers from a particular choice of the density threshold, a procedure that tries to estimate the cluster tree from $\{G_{\varepsilon_0,\sigma}\}_{\sigma\geq 0}$ for a *fixed* ε_0 will be affected by having made fixed choice for the spatial (metric dependent) scale ε_0.

We claim that it may in fact be more informative to encode all possible choices of scale into an *invariant* richer than just a single cluster tree. The invariant we construct out of the family $\{G_{\varepsilon,\sigma}\}_{\varepsilon\geq 0,\sigma\geq 0}$ can be regarded as a generalization of both hierarchical clustering and the cluster tree. In fact, a *slice* of the invariant for a fixed value of ε yields a cluster tree estimate, whereas a slice for a fixed value of σ yields the dendrogram corresponding to applying HC to X^σ, i.e. a single mode analysis snapshot. Our construction therefore takes into account both the linkage parameter ε, and σ: a parameter related to the function f (e.g. density). This is to be regarded as *multiparameter clustering*.

In this paper, we produce a variation of the theme in Carlsson and Mémoli (2008). By first identifying desirable properties of such multi-parameter clustering procedures, we then propose a set of axioms for such methods. We prove a *uniqueness/characterization theorem* (Theorem 1) under these axioms. The procedure singled out by this set of axioms can be regarded as a generalization of both SL HC and the *cluster tree* construction. In addition, in Theorem 2 we establish the precise quantitative (or metric) *stability* of the particular clustering scheme which is characterized by our results.

Our presentation is necessarily concise given the space constraints; more details and elaboration will be presented in a future publication.

2 Notation and Terminology

Let $\mathcal{X}$ denote the collection of all finite metric spaces. Let $\mathcal{X}_1$ be the collection of all finite *filtered metric spaces*, that is triples (X, d_X, f_X) where $(X, d_X) \in \mathcal{X}$ and $f_X : X \to \mathbb{R}$. Given $(X, d_X, f_X) \in \mathcal{X}_1$, for each $\sigma \in \mathbb{R}$ let $X_\sigma = f_X^{-1}((-\infty, \sigma])$. For a finite set X and a symmetric function $W : X \times X \to \mathbb{R}^+$ let $\mathcal{L}(W)$ denote the maximal metric on X less than or equal to W, i.e. $\mathcal{L}(W)(x, x') = \min\left\{\sum_{i=0}^m W(x_i, x_{i+1}) \middle| x = x_0, \ldots, x_m = x', m \in \mathbb{N}\right\}$ for $x, x' \in X$. For a finite metric space (X, d_X), $\mathrm{sep}(X, d_X)$ denotes the minimal distance between any two different points in X. When referring to a metric space (X, d_X) or to a filtered metric space (X, d_X, f_X) we may drop the metric and filter and refer to it by just X. For a topological space S, $\mathcal{B}(S)$ denotes the collection of Borel sets of S. Given a set Z, for a function $h : Z \to \mathbb{R}$, we use the notation $\|h\|_{L_\infty(Z)} = \sup_{z\in Z} |h(z)|$.

3 Two Parameter Hierarchical Clustering: A Characterization Theorem

Definition 1 (Persistent Structures). Given a finite set X, a *persistent structure* on X is a map $Q_X : X \times X \to \mathcal{B}(\mathbb{R}^+ \times \mathbb{R})$ s.t.

1. If $(\varepsilon, \sigma) \in Q_X(x, x')$, then $(\varepsilon + t, \sigma + s) \in Q_X(x, x')$ for all $t, s \geq 0$.
2. If $(\varepsilon_1, \sigma_1) \in Q_X(x, x')$ and $(\varepsilon_2, \sigma_2) \in Q_X(x', x'')$, then $\big(\max(\varepsilon_1, \varepsilon_2), \max(\sigma_1, \sigma_2)\big) \in Q_X(x, x'')$.
3. For all $x, x' \in X$, $\partial Q_X(x, x') \subset Q_X(x, x')$ (technical condition).

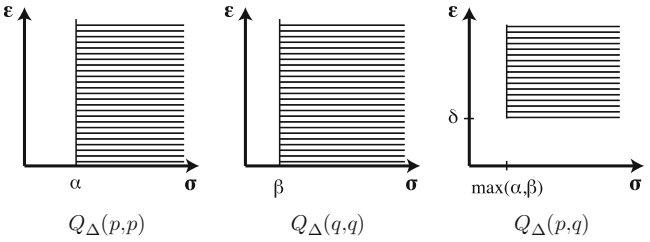


Fig. 1 A simple persistence structure on $\{p, q\}$: Q_Δ

Example 1. Let $\Delta = \{p, q\}$ and Q_Δ be given by the Fig. 1, where $\alpha, \beta, \delta \geq 0$.

Remark 1. Persistent structures are useful constructs for expressing nested associations of points. They can be regarded as a certain generalization of the concept of ultrametrics and therefore of dendrograms (nested families of partitions). In fact, one can see that a persistence structure Q on X gives rise to a family of ultrametrics on X.

We use the language of *categories* and *functors*, see Carlsson and Mémoli (2008) for an exposition relevant to clustering and Mac Lane (1998) for a comprehensive account. Below, $\underline{Sets}$ denotes the category whose objects are sets and whose morphisms are set maps.

Consider the category $\underline{\mathcal{Q}}$ whose objects are pairs (X, Q_X) where X is a finite set and Q_X is a persistent structure on X. Let $\mathcal{Q}$ denote the objects in $\underline{\mathcal{Q}}$. A map $\phi : X \to Y$ is called *persistence preserving* if for all $x, x' \in X$, $Q_X(x, x') \subseteq Q_Y(\phi(x), \phi(x'))$. We declare that $Mor_{\underline{\mathcal{Q}}}((X, Q_X), (Y, Q_Y))$ consists of all persistence preserving maps between X and Y. We define $\underline{\mathcal{M}}^{gen}$ to be the category that has all finite filtered metric spaces as objects, and as morphisms all those maps that are distance non-increasing and filter non-increasing. That is, $\phi \in Mor_{\underline{\mathcal{M}}^{gen}}(X, Y)$ if and only if for all $x, x' \in X$, $d_X(x, x') \geq d_Y(\phi(x), \phi(x'))$ and $f_X(x) \geq f_Y(\phi(x'))$.

In this context, a **clustering functor** will be a functor $\mathcal{C} : \underline{\mathcal{M}}^{gen} \to \underline{\mathcal{Q}}$. Consider the equivalence relation on X_σ given by $x \sim_{(\varepsilon,\sigma)} x'$ if and only if there exists $x_0, \ldots, x_m$ in X s.t. $x_0 = x$, $x_m = x'$, $\max_i d_X(x_i, x_{i+1}) \leq \varepsilon$ and $\max_i f_X(x_i) \leq \sigma$. For $x \in X_\sigma$ let $[x]_{(\varepsilon,\sigma)}$ denote the equivalence class to which x belongs.

Example 2. Consider the functor $\mathcal{C}^* : \underline{\mathcal{M}}^{gen} \to \underline{\mathcal{Q}}$ that when applied to (X, d_X, f_X) produces the object (persistent structure) (X, Q^*_X) where $Q^*_X(x, x') := \{(\varepsilon, \sigma) \in \mathbb{R}^2 | \, x \sim_{(\varepsilon,\sigma)} x'\}$. That $\mathcal{C}^*$ is a functor follows easily from the definitions. The following observations are in order:

- The sets $Q^*_X(x, x')$ are obviously unbounded. They are of the form $\bigcup_{i=1}^{K} [\varepsilon^{(i)}, \infty) \times [\sigma_1^{(i)}, \infty)$. Note that for $x \in X$, $Q^*_X(x, x) = \{(\varepsilon, \sigma) \in \mathbb{R}^2 | \, \varepsilon \geq 0, \sigma \geq f_X(x)\}$.
- Let $B = [x]_{(\varepsilon,\sigma)} \neq [x']_{(\varepsilon,\sigma)} = B'$. Then, clearly, $\min_{x \in B, x' \in B'} d_X(x, x') > \varepsilon$.

- If (ε,σ) are s.t. $\mathrm{sep}(X_\sigma, d_X) > \varepsilon$, then $(\varepsilon,\sigma) \notin Q_X^*(x,x')$ for all x, x' in X_σ with $x \neq x'$. Indeed, otherwise let $x, x', x_0, \ldots, x_n \in X$ be s.t. $x_0 = x$, $x_n = x'$, $d_X(x_i, x_{i+1}) \leq \varepsilon$ and $f_X(x_i) \leq \sigma$. Since $x_i \in X_\sigma$ for $i \in \{0,\ldots,n\}$, and $x \neq x'$, there are at least two different consecutive points in $\{x_0, x_1, \ldots, x_n\}$ whose distance is not greater than ε, a contradiction.
- For $t \geq 0$ let $\sigma_X^t : X \times X \to \mathbb{R}$ be defined by $(x,x') \mapsto \inf\left\{\sigma |\, x \sim_{(t,\sigma)} x'\right\}$. This gives rise to a tree and can be likened to the cluster tree construction of Stuetzle.

We have the following characterization/uniqueness theorem.

Theorem 1. *Let $\mathcal{C} : \underline{\mathcal{M}}^{gen} \to \underline{\mathcal{Q}}$ be a functor which satisfies the following conditions.*

(I) Let $\alpha : \underline{\mathcal{M}}^{gen} \to \underline{Sets}$ and $\beta : \underline{\mathcal{Q}} \to \underline{Sets}$ be the forgetful functors $(X, d_X, f_X) \to X$ and $(X, Q_X) \to X$, which forget the metric and filter, and persistence structure, respectively, and only "remember" the underlying sets X. Then we assume that $\beta \circ \Psi = \alpha$. This means that the underlying set of the persistent structure associated to a metric space is just the underlying set of the metric space.

(II) For $\delta \geq 0$ and $\alpha, \beta \in \mathbb{R}$ let $\Delta(\delta,\alpha,\beta) = (\{p,q\}, \left(\begin{smallmatrix} 0 & \delta \\ \delta & 0 \end{smallmatrix}\right), \{\alpha,\beta\})$ denote the two point filtered metric space with underlying set $\{p,q\}$, where $dist(p,q) = \delta$ and $f_\Delta(p) = \alpha$ and $f_\Delta(q) = \beta$. Then $\mathcal{C}(\Delta(\delta,\alpha,\beta))$ is the persistent structure $(\{p,q\}, Q_\Delta)$ whose underlying set is $\{p,q\}$ and where Q_Δ is given by the construction shown in Fig. 1.

(III) Given $(\varepsilon,\sigma) \in \mathbb{R}^+ \times \mathbb{R}$ and finite filtered metric space (X, d_X, f_X), then $sep(X_\sigma) > \varepsilon$ implies that $(\varepsilon,\sigma) \notin Q_X(x,x')$ for any $x, x' \in X_\sigma$, $x \neq x'$.

Then $\mathcal{C}$ is equal to the functor $\mathcal{C}^*$.

Proof. We sketch the proof. Let (X, d_X, f_X) be a finite filtered metric space. Write $(X, Q_X) = \mathcal{C}(X, d_X, f_X)$. Also, write $(X, Q_X^*) = \mathcal{C}^*(X, d_X, f_X)$.

(1) Let $x, x' \in X$ and $(\varepsilon,\sigma) \in \mathbb{R}^+ \times \mathbb{R}$ be s.t. $(\varepsilon,\sigma) \in Q_X(x,x')$. We will prove that $(\varepsilon,\sigma) \in Q_X^*(x,x')$ as well. Consider the filtered metric space (X', d', f') where $X' = X \setminus \sim_{(\varepsilon,\sigma)}$. Let $\phi : X \to X'$ be given by $x \mapsto [x]_{(\varepsilon,\sigma)}$. For $\alpha, \beta \in X'$ let $W(\alpha,\beta) := \min_{x \in \phi^{-1}(\alpha),\, x' \in \phi^{-1}(\beta)} d_X(x,x')$. Note that by the discussion in Example 2, $\min_{\alpha \neq \beta} W(\alpha,\beta) > \varepsilon$ for $\alpha, \beta \in X'$. Define d' to be the maximal metric pointwisely less than or equal W, i.e. $d' = \mathcal{L}(W)$. Finally , let $f' : X' \to \mathbb{R}$ be given by $\alpha \mapsto \min_{x \in \phi^{-1}(\alpha)} f_X(x)$. Note that by construction, $X'_\sigma = X'$ and $\mathrm{sep}(X', d') > \varepsilon$.

Now, also by construction it holds that $\phi \in Mor_{\underline{\mathcal{M}}^{gen}}(X, X')$. By functoriality we then have $Q_X \subseteq Q_{X'} \circ (\phi,\phi)$, and in particular, we have that $(\varepsilon,\sigma) \in Q_{X'}(\phi(x),\phi(x'))$. Note that we must have $\phi(x) = \phi(x')$ for otherwise, condition (III) together with $\mathrm{sep}(X', d_{X'}) > \varepsilon$ give a contradiction. This means that $[x]_{(\varepsilon,\sigma)} = [x']_{(\varepsilon,\sigma)}$, hence, by definition of $\mathcal{C}^*$, $(\varepsilon,\sigma) \in Q_X^*(x,x')$.

(2) Let $x, x' \in X$ and $(\varepsilon,\sigma) \in \mathbb{R}^+ \times \mathbb{R}$ be s.t. $(\varepsilon,\sigma) \in Q_X^*(x,x')$. Let $x = x_0$, $x_1, \ldots, x_t = x'$ be points in X_σ s.t. $\max_i d_X(x_i, x_{i+1}) \leq \varepsilon$. Fix $i \in \{0,1,\ldots,t-1\}$. Consider the two point filtered metric space $\Delta(\varepsilon,\sigma,\sigma)$ and the map $\psi : \Delta \to X$

given by $\psi(p) = x_i$ and $\psi(q) = x_{i+1}$. Note that by construction $\psi \in Mor_{\underline{\mathcal{M}}^{gen}}(\Delta, X)$. Then, $Q_\Delta \subseteq Q_X \circ (\psi, \psi)$, and in particular (check Fig. 1), $(\varepsilon, \sigma) \in Q_X(x_i, x_{i+1})$. Since i was arbitrary, by applying property 2. in Definition 1 repeatedly, we obtain that $(\varepsilon, \sigma) \in Q_X(x, x')$. This concludes the proof.

Example 3. As a simple practical tool for the analysis of data one could use the following construction: for a given triple (X, d, f) let $K_X : \mathbb{R}^+ \times \mathbb{R} \to \mathbb{N}$ be given by $(\varepsilon, \sigma) \mapsto \#\big(X_\sigma \setminus \sim_{(\varepsilon,\sigma)}\big)$, i.e. the number of equivalence classes of X_σ under $\sim_{(\varepsilon,\sigma)}$.

4 Metric Stability of $\mathcal{C}^*$

For sets A and B, a subset $R \subset A \times B$ is a *correspondence* (between A and B) if and only if (1) $\forall\, a \in A$, there exists $b \in B$ s.t. $(a, b) \in R$; and (2) $\forall\, b \in B$, there exists $a \in X$ s.t. $(a, b) \in R$. Let $\mathcal{R}(A, B)$ denote the set of all possible correspondences between sets A and B.

Consider compact metric spaces (X, d_X) and (Y, d_Y). Let $\Gamma_{X,Y} : X \times Y \times X \times Y \to \mathbb{R}^+$ be given by $(x, y, x', y') \mapsto |d_X(x, x') - d_Y(y, y')|$. Then, the **Gromov-Hausdorff distance** (Burago 2001) between X and Y is given by $d_{\mathcal{GH}}(X, Y) := \inf_{R \in \mathcal{R}(X,Y)} \|\Gamma_{X,Y}\|_{L^\infty(R \times R)}$. The Gromov-Hausdorff distance is a metric on the collection of all isometry classes of compact metric spaces (Burago 2001). We modify the expression of the Gromov-Hausdorff distance in order to define a metric for filtered metric spaces. We deem two spaces X, Y in $\mathcal{X}_1$ *isomorphic* whenever there exists an isometry $\Psi : (X, d_X) \to (Y, d_Y)$ such that $f(x) = g \circ \Psi(x)$ for all $x \in X$.

Definition 2. Let $\mathbf{D} : \mathcal{X}_1 \times \mathcal{X}_1 \to \mathbb{R}^+$ be given by

$$\mathbf{D}(X, Y) := \min_{R \in \mathcal{R}(X,Y)} \max\left(\|\Gamma_{X,Y}\|_{L^\infty(R \times R)}, \|f_X - f_Y\|_{L^\infty(R)}\right), \ X, Y \in \mathcal{X}_1.$$

Proposition 1. *The function* $\mathbf{D}$ *defined above is a metric on (the set of isomorphism classes of)* $\mathcal{X}_1$.

We say that two persistent structures (X, Q_X) and (Y, Q_Y) are *isomorphic* and write $(X, Q_X) \simeq (Y, Q_Y)$, if and only if there exist a bijection $\Phi : X \to Y$ s.t. $Q_Y = Q_X \circ (\Phi, \Phi)$. We define a metric on the collection $\mathcal{Q}$ of all persistent structures by

$$d_{\mathcal{Q}}(X, Y) := \min_{R \in \mathcal{R}(X,Y)} \max_{(x,y),(x',y') \in R} d_{\mathcal{H}}^{(\mathbb{R}^2, L^\infty)}\big(Q_X(x, x'), Q_Y(y, y')\big) \qquad (1)$$

In (1) above, $d_{\mathcal{H}}^{(\mathbb{R}^2, L^\infty)}$ stands for the *Hausdorff distance* (Burago 2001) on subsets of the plane under the L^∞ metric.

Proposition 2. *$d_{\mathcal{Q}}$ defines a metric on (the isomorphism classes of) $\mathcal{Q}$.*

Now one has *stability* on the functor $\mathcal{C}^*$, i.e. the application $(X, d_X, f_X) \mapsto (X, Q^*_X)$ is stable in an appropriate sense.

Theorem 2. *For two filtered spaces (X, d_X, f_X) and (Y, d_Y, f_Y) in $\mathcal{X}_1$ consider the associated persistent structures (X, Q^*_X) and (Y, Q^*_Y) defined in Example 2. Then, one has $d_{\mathcal{Q}}((X, Q^*_X), (Y, Q^*_Y)) \leq \mathbf{D}(X, Y)$.*

Acknowledgements Our research has been supported by DARPA grant number HR0011-05-1-0007 (GC and FM), NSF DMS-0406992 (GC) and ONR grant number N00014-09-1-0783 (FM).

References

Anthony Wong, M., & Lane, T. (1983). A kth nearest neighbour clustering procedure. *Journal of the Royal Statistical Society: Series B, 45*(3), 362–368.

Ben-David, S., von Luxburg, U., & Pál, D. (2006). A sober look at clustering stability. In G. Lugosi & H.-U. Simon (Eds.), *COLT*, volume 4005 of *Lecture Notes in Computer Science* (pp. 5–19). Berlin, Heidelberg, New York: Springer.

Biau, G., Cadre, B., & Pelletier, B. (2007). A graph-based estimator of the number of clusters. *ESAIM Probability and Statistics, 11*, 272–280.

Burago, D., Burago, Y., & Ivanov, S. (2001). *A course in metric geometry*, volume 33 of *AMS Graduate Studies in Maths*. American Mathematical Society.

Carlsson, G., & Mémoli, F. (2008). Persistent clustering and a theorem of J. Kleinberg. *ArXiv e-prints*.

Cuevas, A., Febrero, M., & Fraiman, R. (2001). Cluster analysis: A further approach based on density estimation. *Computational Statistics and Data Analysis, 36*(4), 441–459.

Edelsbrunner, H., Letscher, D., & Zomorodian, A. (2000). Topological persistence and simplification. In *Proceedings of the 41st Annual IEEE Symposium Foundation of Computer Science* (pp. 454–463).

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise*, 226–231. Menlo Park, CA, USA: AAAI Press.

Hartigan, J. A. (1975). *Clustering algorithms*. New York-London-Sydney: Wiley. Wiley Series in Probability and Mathematical Statistics.

Hartigan, J. A. (1981). Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association, 76*(374), 388–394.

Jardine, N., & Sibson, R. (1971). *Mathematical taxonomy*. London: Wiley. Wiley Series in Probability and Mathematical Statistics.

Kleinberg, J. M. (2002). An impossibility theorem for clustering. In S. Becker, S. Thrun and K. Obermayer (Eds.), *NIPS* (pp. 446–453). Cambridge, MA: MIT Press.

Klemelä, J. (2004). Visualization of multivariate density estimates with level set trees. *Journal of Computational and Graphical Statistics, 13*(3), 599–620.

Lance, G. N., & Williams, W. T. (1967). A general theory of classificatory sorting strategies 1. Hierarchical systems. *Computer Journal, 9*(4), 373–380.

Mac Lane, S. (1998). *Categories for the working mathematician* (2nd ed.), Vol. 5 of *Graduate Texts in Mathematics*. New York: Springer-Verlag.

Stuetzle, W. (2003). Estimating the cluster type of a density by analyzing the minimal spanning tree of a sample. *Journal of Classification, 20*(1), 25–47.

Stuetzle, W., & Nugent, R. (2008). *A generalized single linkage method for estimating the cluster tree of a density*.

Wishart, D. (1969). Mode analysis: a generalization of nearest neighbor which reduces chaining effects. In *Numerical Taxonomy* (pp. 282–311). London: Academic Press.

Unsupervised Sparsification of Similarity Graphs

Tim Gollub and Benno Stein

Abstract Cluster analysis often grapples with high-dimensional and noisy data. The paper in hand identifies sparsification as an approach to address this problem. Sparsification improves both the runtime and the quality of cluster algorithms that exploit pairwise object similarities, i.e., that rely on similarity graphs. Sparsification has been addressed in the field of graphical cluster algorithms in the past, but the developed approaches leave the burden of parameter tuning to the user. Our approach to sparsification relies on the inherent characteristics of the data and is completely unsupervised. It leads to significant improvements in the cluster quality and outperforms even the optimum supervised approaches to sparsification that rely on a single global threshold.

1 Introduction and Related Work

Cluster analysis deals with the problem of finding natural groups in large sets of data. Extensive discourses on clustering techniques are given in Everitt (1993), Jain et al. (2000), Kaufman and Rousseuw (1990), Stein and Meyer zu Eißen (2003). For the purpose of this paper it is sufficient to distinguish between clustering techniques that are based on a similarity graph versus techniques that are exemplar-based. The contribution of our research is to the former class of algorithms. Figure 1 provides an overview of algorithms that are based on similarity graphs.

To motivate sparsification as a vital part of cluster analysis, consider the conceptual model of a cluster analysis process shown in Fig. 2. The similarity graph G of a set of objects $O = \{o_1, o_2, \ldots, o_m\}$ is derived by estimating the similarities between all pairs of objects. Similarities between real-world objects such as documents cannot be assessed directly (unless done by human) but require a model formation or feature extraction step, resulting in a set of object *representations*

T. Gollub (✉)
Faculty of Media/Media Systems, Bauhaus-Universität Weimar, Germany
e-mail: Tim.Gollub@uni-weimar.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_7,

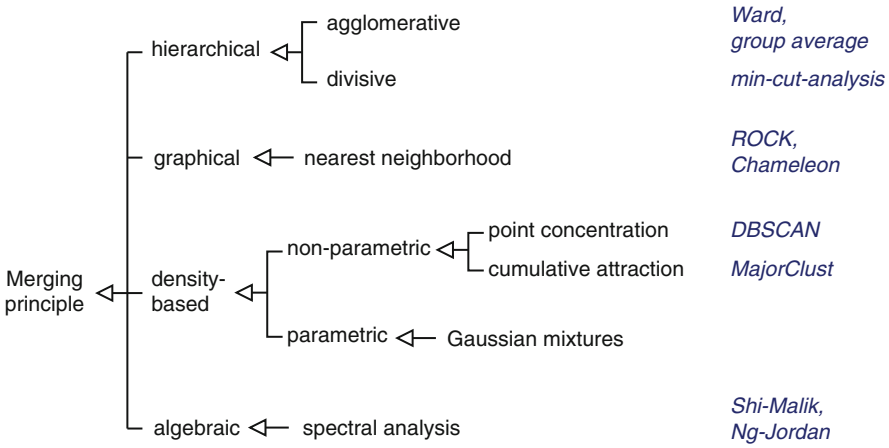


Fig. 1 Taxonomy of cluster algorithms using similarity graphs

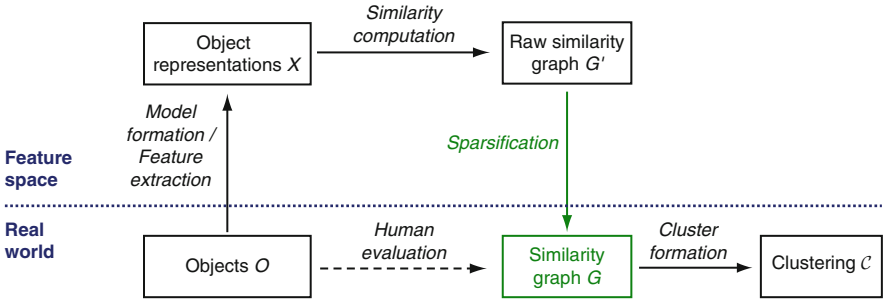


Fig. 2 Cluster analysis. A four-step conceptual model

$X = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_m}\}$. A vector $\mathbf{x_i} \in X$ corresponds to n features of an object o_i and comprises the respective feature weights, i.e., $\mathbf{x_i} = (w_{i_1}, w_{i_2}, \ldots, w_{i_n})^T$. A similarity function $s(\mathbf{x_i}, \mathbf{x_j}) \rightarrow [0, 1]$ is applied to all pairs in X to construct the raw similarity graph G'. If the model formation step is adequate,[1] G' resembles the similarity graph G of the real-world objects O. However, the indirection shown in the upper part of Fig. 2 introduces undesired imprecision. Among others, imprecision is introduced in the course of feature selection, feature computation, or similarity quantification. Hence the raw similarity graph G' models the similarities between the real-world objects O only approximately. Note that a cluster algorithm takes the similarity scores in G' at face value and runs the risk to make wrong decisions, especially in tie situations. Here sparsification comes into play. By modifying the raw similarity graph G', a smart sparsification obtains a more veritable similarity graph G.

[1] In the sense of Minsky (1965): $\mathbf{x_i}$ can answer the interesting question about o_i.

In Kumar (2000), Luxburg (2007), Kumar and Luxburg report on two major approaches to sparsification. The first one uses a global threshold τ to eliminate all edges with a similarity score below this value. As will be discussed in greater detail in Chap. 2, this approach has its major drawback in disregarding regions of variable density in the object space. The second approach to sparsification is more sensible. It discards all edges of G' that are not among the k strongest edges of a node. Several variants of this nearest neighbor sparsification are discussed in Ertöz et al. (2003), Guha et al. (1999), Karypis et al. (1999). While nearest neighbor sparsification longs for different regions in the document space, it comes at the price that the parameter k is application-dependent and has to be chosen carefully.

Our new approach to sparsification adapts itself in an unsupervised manner. It computes an expected similarity score for every edge in the graph G', and only similarity scores surpassing this expectation remain in the thinned-out graph G. To examine the potential of our idea, we conduct two experiments on four collections of text documents. In the first experiment the accuracy of our sparsification technique is compared to the best performing approach that uses a global threshold. In the second experiment raw and thinned-out similarity graphs are analyzed by a density-based cluster algorithm in order to evaluate the gain in clustering quality achieved by sparsification. The results of our experiments are promising in every respect: sparsification increases the quality of the clusterings. Even more, our approach excels in every experiment even the optimum sparsification that relies on a global threshold.

The organization of the paper is as follows. Section 2 gives a definition of sparsification in the context of cluster analysis and presents the new unsupervised sparsification approach. Section 3 reports on the experiments.

2 Sparsification

In the field of computational theory, sparsification is understood as a technique to guarantee a desired time bound when designing dynamic algorithms (Black 2004). In cluster analysis research, improving the efficiency of an approach is of interest as well,[2] but sparsification is also used to enhance the cluster *quality*. Kumar (2000) states the goal of sparsification as the *"efficient and effective identification of the core points belonging to clusters and subclusters"*. This definition, though reasonable, is closely related to the author's approach to graphical clustering. Here we propose a more general definition in the context of cluster analysis:

> *Sparsification is the interpretation of the similarity scores in the feature space in order to enhance the quality and the effort of the cluster formation task.*

Ideally, sparsification sets the similarity scores of edges between two clusters (inter-class edges) to zero, while setting the edge scores within clusters (intra-class edges)

[2] E.g., spectral clustering is efficient only with sparse matrices (Luxburg 2007).

to 1. Let $c(o_i) \rightarrow \{1, \ldots, l\}$ assign the true class label to each object $o \in O$. Then, the optimum sparse similarity graph G fulfills the following condition:

$$\varphi(o_i, o_j) = \begin{cases} 1, & \text{if } c(o_i) = c(o_j) \\ 0 & \text{otherwise,} \end{cases}$$

where $\varphi(o_i, o_j)$ denotes the similarity between two real-world objects. The optimum similarity graph is only of theoretical interest since it requires unavailable knowledge about the true class labels. Existing approaches to sparsification work out a notion of probability that two objects belong to the same class. The underlying principle is the nearest neighbor principle. It states that if an object representation $\mathbf{x_1}$ is more similar to a representation $\mathbf{x_2}$ than to another representation $\mathbf{x_3}$, then the probability that $\mathbf{x_2}$ belongs to the same class as $\mathbf{x_1}$ should be higher than the probability that $\mathbf{x_3}$ belongs to the same class as $\mathbf{x_1}$:

$$s(\mathbf{x_1}, \mathbf{x_2}) > s(\mathbf{x_1}, \mathbf{x_3}) \quad \Leftrightarrow \quad P(c(o_1) = c(o_2)) > P(c(o_1) = c(o_3))$$

Upon this supposition several approaches, including ours, have been suggested.

2.1 Existing Approaches

The most common approach to sparsification is the use of a global threshold τ. Every similarity score below the threshold is discarded from the similarity graph. While this approach can be applied efficiently it has two serious drawbacks. First, the threshold's optimum value varies under different sets of objects and has to be found empirically. Second, applying one global threshold does not account for different regions in the object space. As illustrated in Fig. 3 one has to cope with clusters where objects are connected much loser compared to other clusters. In such a situation the upper bound for the threshold is determined by the cluster of the lowest density.

The second approach to sparsification relies on the construction of a k-nearest neighbor graph of G'. The k-nearest neighbor graph retains those edges which are among the heaviest k edges of a node (= link to the k nearest neighbors). Several variants of this algorithm exist. The mutual k-nearest neighbor graph is constructed by discarding each edge for which the incident nodes are not among the k nearest neighbors of each other. Another interesting variant is called shared nearest neighbor graph, where the edges of an ordinary k-nearest neighbor graph are weighted according to the number of neighbors the incident nodes have in common. As illustrated in Fig. 3 a k-nearest neighbor graph is able to retain regions of different density in the object space. The main problem is the proper adjustment of the parameter k. And, since the optimum k heavily depends on the (unknown)

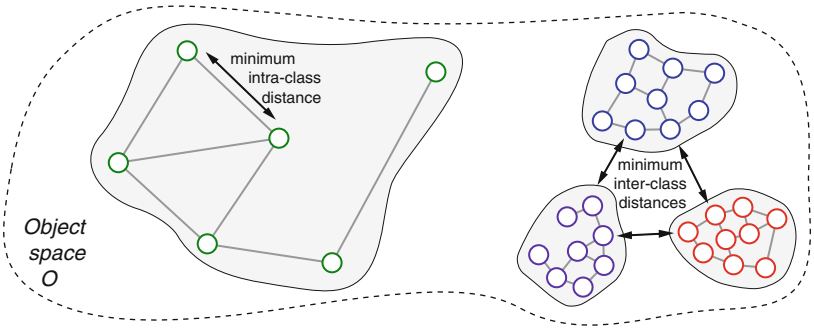


Fig. 3 Two different regions in the object space. The dense clusters on the right-hand side could be thinned-out effectively by applying a threshold reflecting the minimum inter-class distance. Within the cluster on the left, however, this threshold would eliminate all intra-class edges. A convincing result is obtained by constructing the mutual three-nearest neighbor graph (illustrated by the indicated edges). Every node has at least one intra-class edge; all inter-class edges are discarded

number and size of the classes, even finding a limiting range of promising choices is difficult. Ertöz et al. (2003) state:

> *"The neighborhood list size, k, is the most important parameter as it determines the granularity of the clusters. If k is too small, even a uniform cluster will be broken up into pieces due to the local variations in the similarity [...]. On the other hand, if k is too large, then the algorithm will tend to find only a few large, well-separated clusters, and small local variations in similarity will not have an impact."*

Hence, the construction of a suitable sparse similarity graph requires the generation and evaluation of a large number of candidates. Note that – more than runtime – the identification of a sensible internal evaluation measure is the limiting factor in this connection.

2.2 An Object-specific, Unsupervised Approach to Sparsification

Our goal is to provide a completely unsupervised approach to sparsification, while striving for the performance of the existing supervised approaches. To achieve this we claim that two objects in the thinned-out graph G are only allowed to share an edge, if the probability that they belong to the same cluster is high. In particular we propose that the following relation must hold:

$$P(c(o_1) = c(o_2)) \;>\; \max\{P(c(o_1) = c(o_{rand})),\; P(c(o_2) = c(o_{rand}))\},$$

with $o_{rand} \in O \setminus \{o_1, o_2\}$. I.e., the probability that two objects, o_1 and o_2, belong to the same cluster must exceed the probabilities that some randomly drawn object from O belongs to the same cluster as o_1 or o_2. Given this postulation, the nearest

neighbor principle is used to establish a relation concerning the similarity scores of the corresponding object representations:

$$s(\mathbf{x_1}, \mathbf{x_2}) \; > \; \max\{s(\mathbf{x_1}, \overline{\mathbf{x}}),\; s(\mathbf{x_2}, \overline{\mathbf{x}})\},$$

where $\overline{\mathbf{x}}$ is a virtual object representation reflecting the characteristics of the object set. It comprises the average weights of all object representations in X:

$$\overline{\mathbf{x}} = (\overline{w}_1, \ldots, \overline{w}_n)^T \quad \text{with } \overline{w}_j = \frac{\sum_{i=0}^{m} w_{i,j}}{m}.$$

If the similarity score of two object representations does not exceed the postulated score, the respective edge is classified as an inter-class edge and is discarded. Altogether, the decision rule $\overline{\varphi}$ for unsupervised sparsification reads as follows:

$$\overline{\varphi}(o_1, o_2) := \begin{cases} s(\mathbf{x_1}, \mathbf{x_2}), & \text{if } s(\mathbf{x_1}, \mathbf{x_2}) \; > \; \max\{s(\mathbf{x_1}, \overline{\mathbf{x}}),\; s(\mathbf{x_2}, \overline{\mathbf{x}})\} \\ 0 & \text{otherwise.} \end{cases}$$

The decision rule above yields convincing results in our sparsification experiments. Nevertheless, cluster algorithms that are extremely sensitive to noise benefit from a more exhaustive sparsification. To account for this, the notion of *significance* is introduced into the formula by modifying the virtual object representation $\overline{\mathbf{x}}$. In the following formula the maximum weight of each feature w_i^* is considered as an upper bound, and the harmonic mean between this bound and the averaged feature weight is computed:

$$\widehat{\mathbf{x}} = (\widehat{w}_1, \ldots, \widehat{w}_n)^T \quad \text{with } \widehat{w}_i = \frac{2 \cdot w_i^* \cdot \overline{w_i}}{w_i^* + \overline{w_i}}.$$

The corresponding stricter decision rule $\widehat{\varphi}$, which accounts for significance, is derived by substituting $\widehat{\mathbf{x}}$ for $\overline{\mathbf{x}}$ in the decision rule $\overline{\varphi}$.

3 Evaluation

To evaluate the performance of our unsupervised approach to sparsification, four test collections were constructed from the Reuters news corpus RCV1 (Rose et al. 2002). The collections vary with respect to the number of documents, the number of categories, as well as by the way the documents are distributed across the classes (cf. Table 1).

The documents are represented using the vector space model with normalized *tf*-feature-weights (Salton et al. 1975), having applied Porter stemming and stopword elimination. The similarity between two documents is computed as the dot product

Table 1 Properties of the four test collections. Based on the first collection, one attribute at a time is altered in the subsequent collections

Collection	Categories	Documents	Distribution
1	4	10.000	random
2	4	10.000	uniform
3	4	2.000	random
4	10	10.000	random

Table 2 Averaged results of the experimental analysis. The first and the second row show the results with the minimum and the optimum global threshold respectively. The third and the fourth row report on our unsupervised approach, employing the virtual documents $\overline{\mathbf{x}}$ and $\widehat{\mathbf{x}}$. Column 4 reports on the F-measure in the first experiment (sparsification task), the rightmost column reports on the quality of the clusterings produced by MajorClust

Approach	% of retained intra-class edges	% of discarded inter-class edges	F-measure (sparsification)	F-measure (clustering)
$\tau = min$	100.0%	6.0%	0.43	0.23
$\tau^* = 0.075$	59.9%	83.0%	0.59	0.61
$\overline{\varphi}$	66.8%	85.9%	**0.63**	0.68
$\widehat{\varphi}$	37.4%	97.4%	0.50	**0.76**

of their representations. The nonzero similarity scores are manually divided into intra-class and inter-class scores.

In the first experiment we are interested in the accuracy of our approach. It is specified in terms of the F-measure, $F = \frac{2 \times precision \times recall}{precision + recall}$. While *precision* denotes the proportion of intra-class edges in the thinned-out graph, *recall* is determined by the proportion of intra-class edges retained. The global threshold sparsification that classifies the edges best (= highest F-measure) is identified by an exhaustive search and is compared to the results obtained by our unsupervised approach. The average results of the experiment are shown in Column 4 of Table 2. Our unsupervised sparsification approach with the virtual object $\overline{\mathbf{x}}$ (Row 3) outperforms sparsification with the optimum global threshold (Row 2).

In the second experiment the thinned-out similarity graphs are given to MajorClust, a representative of the density-based cluster formation paradigm (cf. Fig. 1). Here we use the classification-oriented F-measure, described, e.g., in Rosenberg and Hirschberg (2007), to determine the quality of the resulting clusterings. The average results are shown in the rightmost column of Table 2. The first row serves as a baseline: these values are achieved by applying the maximum global threshold that retains 100% of the intra-class edges. Note that sparsification in general raises the cluster quality. Comparing the different approaches to sparsification, our unsupervised approach with the virtual object $\overline{\mathbf{x}}$ again outperforms the global threshold sparsification. Interestingly, sparsification with the virtual object $\widehat{\mathbf{x}}$, which retains only 37.4% of the intra-class edges but discards 97.4% of the inter-class edges, attains the highest cluster qualities (Row 4).

4 Conclusion

The main contribution of this paper is a new, unsupervised approach to sparsification. We argue that existing cluster analysis technology is over-strained with the amount of noise that is typical for most categorization and classification tasks, e.g., in information retrieval. A preprocessing of the similarity graph in the form of a sparsification step considerably improves the cluster performance.

The outstanding property of the proposed rule is the consideration of the specific similarity distributions within the set of objects, while being parameterless at the same time. Our analysis shows that even sparsification with the optimum global threshold is outperformed. Recall in this context that a comparison to the optimum threshold is only of theoretical interest: in practical applications, cluster analysis happens unsupervised, and the optimum threshold is not at hand. This fact underlines the impact of the proposed strategy.

A still unanswered research question is the performance of our approach in comparison to a k-nearest neighbor approach. A preliminary evaluation of smaller document sets (up to 2,000 documents) revealed, that our unsupervised approach to sparsification is as effective as the best performing mutual k-nearest neighbor graph in 86% of 126 different cases (Gollub 2008).

References

Black, P. E. (2004). "Sparsification", in dictionary of algorithms and data structures [online]. In U.S. National Institute of Standards and Technology, (Eds.), *Algorithms and theory of computation handbook*. Boca Raton: CRC Press LLC. URL http://www.itl.nist.gov/div897/sqg/dads/HTML/sparsificatn.html.

Ertöz, L., Steinbach, M., & Kumar, V. (2003). Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *SDM*.

Everitt, B. S. (1993). Cluster analysis. New York: Toronto.

Gollub, T. (2008). Verfahren zur modellbildung für das dokumenten-clustering. Diplomarbeit, Bauhaus-Universität Weimar, Fakultät Medien, Mediensysteme, April 2008. In German.

Guha, S., Rastogi, R., & Shim, K. (1999). Rock: A robust clustering algorithm for categorical attributes. In *ICDE '99: Proceedings of the 15th International Conference on Data Engineering* (p. 512). Washington, DC, USA: IEEE Computer Society. ISBN 0-7695-0071-4.

Jain, A. K., Murty, M. N., & Flynn, P. J. (2000). Data clustering: A review. *ACM Computing Surveys (CSUR), 31*(3), 264–323. ISSN 0360-0300. http://doi.acm.org/10.1145/331499.331504.

Karypis, G., Han, E.-H., & Kumar, V. (1999). Chameleon: A hierarchical clustering algorithm using dynamic modeling. Technical Report Paper No. 432, Minneapolis: University of Minnesota.

Kaufman, L., & Rousseuw, P. J. (1990). *Finding groups in data*. New York: Wiley.

Kumar, V. (2000). An introduction to cluster analysis for data mining. Technical report, CS Dept, University of Minnesota, USA.

Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing, 17*(4), 395–416. ISSN 0960-3174. http://dx.doi.org/10.1007/s11222-007-9033-z.

Minsky, M. (1965). Models, minds, machines. In *Proceedings of the IFIP Congress* (pp. 45–49).

Rose, T. G., Stevenson, M., & Whitehead, M. (2002). The reuters corpus volume 1 – From yesterday's news to tomorrow's language resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation*.

Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in natural language processing and computational natural language learning (EMNLP-CoNLL)* (pp. 410–420).

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communication ACM, 18*(11), 613–620.

Stein, B., & Meyer zu Eißen, S. (2003). Automatic document categorization: interpreting the perfomance of clustering algorithms. In A. Günter, R. Kruse & B. Neumann (Eds.), *KI 2003: Advances in artificial intelligence*, volume 2821 LNAI of *Lecture Notes in Artificial Intelligence* (pp. 254–266). Springer, September 2003. ISBN 3-540-20059-2.

Simultaneous Clustering and Dimensionality Reduction Using Variational Bayesian Mixture Model

Kazuho Watanabe, Shotaro Akaho, Shinichiro Omachi, and Masato Okada

Abstract Exponential principal component analysis (e-PCA) provides a framework for appropriately dealing with various data types such as binary and integer for which the Gaussian assumption on the data distribution is inappropriate. In this paper, we develop a simultaneous dimensionality reduction and clustering technique based on a latent variable model for the e-PCA. Assuming the discrete distribution on the latent variable leads to mixture models with constraint on their parameters. We derive a learning algorithm for those mixture models based on the variational Bayes method. Although intractable integration is required to implement the algorithm, an approximation technique using Laplace's method allows us to carry out clustering on an arbitrary subspace. Numerical experiments on handwritten digits data demonstrate its effectiveness for extracting the structures of data as a visualization technique and its high generalization ability as a density estimation model.

1 Introduction

Exponential principal component analysis (e-PCA) has been proposed as a dimensionality reduction method that extracts a low dimensional subspace in the space of probability distributions (Akaho 2004; Collins et al. 2002). This method provides a framework for introducing appropriate distance measures for special data types such as binary and integer. The original principal component analysis (PCA) uses the squared Euclidean distance based on the Gaussian assumption of the data. Extending this assumption to the exponential family distribution introduces more appropriate distance measures for various data types (Collins et al. 2002). This also resolves the problem that the projections of data obtained by the original PCA can be outside the domain. Such extension is not restricted only to dimension reduction, but it can be applied to clustering as well (Banerjee et al. 2005).

K. Watanabe (✉)
Nara Institute of Science and Technology, 8916-5, Takayama-cho, Ikoma, Nara, 630-0192, Japan
e-mail: wkazuho@is.naist.jp

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_8,

One drawback for the e-PCA is that it is not associated with a statistical estimation model. Several advantages have been provided by interpreting dimensionality reduction and clustering methods as estimations in statistical models, including the application of the Bayesian framework and the use of the semi-supervised approach. Among them, the Bayesian framework offers mechanisms for selecting the intrinsic dimensionality in dimension reduction methods and the number of clusters (populations) in clustering methods. In fact, the probabilistic PCA (Tipping and Bishop 1999), proposed as a latent variable model for the PCA, became the basis for several methods such as the Bayesian PCA (Bishop 1999).

In this paper, we introduce a latent variable model for the e-PCA and present a simultaneous clustering and dimensionality reduction technique using the mixture model whose parameter vectors are constrained to a low dimensional subspace of exponential family distributions. As the main contribution of this paper, we derive a learning algorithm for the constrained mixture model based on the variational Bayes method. We consider the variational Bayesian algorithm for a fixed subspace. However, it requires intractable integration. We apply a Laplace approximation method to it to achieve clustering on an arbitrary subspace. Furthermore, devising methods to estimate the basis vectors of the subspace and the projections of the data and combining them with the above clustering algorithm, we complete a framework for simultaneous clustering and dimensionality reduction. The variational Bayesian algorithm naturally provides a criterion for selecting the intrinsic dimensionality as well as the number of clusters present.

We demonstrate the properties of the derived algorithm by the experiment conducted for the handwritten digits data with discrete features. By reducing the dimensionality to two or three, our framework also provides a visualization technique. Comparisons of the derived method to the conventional methods such as Gaussian mixture modelling show that it well preserves multi-modality of the high dimensional distribution in the visualization space and it has high generalization ability as a density estimation model.

2 Exponential Family and e-PCA

A statistical model $p(\mathbf{x}|\theta)$ of a random vector $\mathbf{x} \in \Xi$ is called the exponential family if its probability density function (or probability function) has the following form,

$$p(\mathbf{x}|\theta) = \exp\{\theta \cdot \mathbf{F}(\mathbf{x}) + F_0(\mathbf{x}) - G(\theta)\}, \tag{1}$$

where $\theta = (\theta_1, \ldots, \theta_M)^T \in \Theta$ is called the natural parameter and $\theta \cdot \mathbf{F}(\mathbf{x}) = \sum_{j=1}^{M} \theta_j F_j(\mathbf{x})$ is the inner product of θ and $\mathbf{F}(\mathbf{x}) = (F_1(\mathbf{x}), \ldots, F_M(\mathbf{x}))^T$. The function $F_0(\mathbf{x})$ is real-valued and $G(\theta)$ ensures that $p(\mathbf{x}|\theta)$ is a probability density function. The expectation parameter is defined by $\eta = (\eta_1, \eta_2, \ldots, \eta_M)$ where $\eta_i = E[F_i(\mathbf{x})] = \int F_i(\mathbf{x}) p(\mathbf{x}|\theta) d\mathbf{x}$ is the expectation of $F_i(\mathbf{x})$ with respect to the distribution (1). The exponential family has an important property that there is a

bijection between θ and η given by $\eta = \frac{\partial G(\theta)}{\partial \theta}$. Hence we denote them by $\theta(\eta)$, $\eta(\theta)$ as functions of respective parameters.

The e-PCA is a method to extract a low dimensional subspace in the set of the exponential family distributions. Let

$$\tilde{\theta}(\mathbf{w}) = \mathbf{U}\mathbf{w} + \mathbf{u}_0 = \sum_{j=1}^{L} w_j \mathbf{u}_j + \mathbf{u}_0 \tag{2}$$

represent a point on the L-dimensional subspace in Θ. Here $\mathbf{U}$ is the matrix whose columns are the M-dimensional basis vectors, $\mathbf{u}_1, \ldots, \mathbf{u}_L$ and $\mathbf{w} = (w_1, w_2, \ldots, w_L) \in R^L$ is the low dimensional representation of $\tilde{\theta}$. The set $\{\tilde{\theta}(\mathbf{w}); \mathbf{w} \in R^L\}$ forms the so-called "e-flat" subspace (Akaho 2004; Amari and Nagaoka 2000). It is known that for a point $\theta \in \Theta$, there is a unique point $\tilde{\theta}$ in the e-flat subspace that minimizes the Kullback information $K(\theta||\tilde{\theta})$.[1] Based on this fact, given the samples of the natural parameter $\Theta^n = \{\theta^{(i)}\}_{i=1}^n$ as the training data, the e-PCA searches the latent variables $\mathbf{W}^n = \{\mathbf{w}^{(i)}\}_{i=1}^n$ (also the basis $\mathbf{U}$ and $\mathbf{u}_0$) so as to minimize the objective function, $\sum_{i=1}^n K(\theta^{(i)}||\tilde{\theta}(\mathbf{w}^{(i)}))$. An alternating optimization procedure was derived for the e-PCA (Akaho 2004).

3 Constrained Mixture Model

The probabilistic PCA assumes the joint probability distribution of the data vector $\mathbf{x}$ and the latent variable $\mathbf{w} = (w_1, w_2, \ldots, w_L) \in R^L$ as follows,

$$p(\mathbf{x}, \mathbf{w}) = p(\mathbf{x}|\tilde{\theta}(\mathbf{w}))p(\mathbf{w}), \tag{4}$$

where $\tilde{\theta}(\mathbf{w})$ is defined by (2) (Tipping and Bishop 1999). Throughout this paper, we suppose that $p(\mathbf{x}|\theta)$ in (4) is an exponential family distribution given in (1). The subsequent discussion before Sect. 5 assumes that $\mathbf{U}$ and $\mathbf{u}_0$ are fixed.

Suppose n training samples $X^n = \{\mathbf{x}^{(i)}\}_{i=1}^n$ are given. If we consider $\mathbf{F}(\mathbf{x}^{(i)})$ as a sample of the expectation parameters and $\theta^{(i)}$ as that of the corresponding natural parameters, that is,

$$\mathbf{F}(\mathbf{x}^{(i)}) = \eta(\theta^{(i)}), \tag{5}$$

then the maximum a posteriori estimator of $\mathbf{W}^n$, induced by the above model, corresponds to the e-PCA solution when the latent distribution $p(\mathbf{w}^{(i)})$ is uniform (Watanabe et al. 2009).

[1] The Kullback information between the two distributions, $p(\mathbf{x}|\theta)$ and $p(\mathbf{x}|\tilde{\theta})$ is defined by $K(\theta||\tilde{\theta}) = \int p(\mathbf{x}|\theta) \log \frac{p(\mathbf{x}|\theta)}{p(\mathbf{x}|\tilde{\theta})} d\mathbf{x}$. In the case of the exponential family, it is given by

$$K(\theta||\tilde{\theta}) = (\theta - \tilde{\theta})\eta(\theta) - G(\theta) + G(\tilde{\theta}). \tag{3}$$

To model multiple populations (clusters), we assume a discrete prior distribution. Let δ be the delta function and $\mathbf{a} = \{a_k\}_{k=1}^K$ be a set of real numbers that satisfy $a_k \geq 0$ and $\sum_{k=1}^K a_k = 1$. Assuming the density $p(\mathbf{w})$ of the latent variable to be the discrete distribution $\sum_{k=1}^K a_k \delta(\mathbf{w} - \mathbf{w}_k)$ and marginalizing $\mathbf{w}$ in (4), yield a finite mixture model,

$$p(\mathbf{x}|\omega) = \sum_{k=1}^{K} a_k \, p(\mathbf{x}|\tilde{\theta}(\mathbf{w}_k)). \tag{6}$$

Here $\omega = \{\mathbf{a}, \{\mathbf{w}_k\}_{k=1}^K; \mathbf{w}_k \in R^L\}$ is the parameter of the mixture model and K is the number of components. Each component $p(\mathbf{x}|\tilde{\theta}(\mathbf{w}_k))$ is the exponential family distribution (1) whose parameter $\tilde{\theta}(\mathbf{w}_k)$ defined by (2) is constrained to the L-dimensional subspace. This mixture model was derived based on non-parametric maximum likelihood (ML) estimation of the latent distribution $p(\mathbf{w})$ and an expectation maximization (EM) algorithm was given for parameter estimation (Sajama and Orlitsky 2004).

We take the conjugate prior distribution of the parameter as $p(\omega) = p(\mathbf{a}) \prod_{k=1}^K p(\mathbf{w}_k)$, in which the prior $p(\mathbf{a})$ and the prior $p(\mathbf{w}_k)$ are the Dirichlet and exponential family distributions,

$$p(\mathbf{a}) = \frac{\Gamma(K\phi_0)}{\Gamma(\phi_0)^K} \prod_{k=1}^{K} a_k^{\phi_0 - 1}, \quad p(\mathbf{w}_k) = \exp\{\xi_0(\mathbf{w}_k \cdot \alpha_0 - G(\tilde{\theta}(\mathbf{w}_k))) - \Phi(\alpha_0, \xi_0)\}, \tag{7}$$

with hyperparameters $\phi_0 > 0$, $\mathbf{U}$, $\mathbf{u}_0$, $\alpha_0 \in R^L$ and $\xi_0 > 0$. The function $\Phi(\alpha, \xi)$ of $\xi \in R$ and $\alpha \in R^L$ is defined by

$$\Phi(\alpha, \xi) = \log \int \exp\{\xi(\alpha \cdot \mathbf{w} - G(\tilde{\theta}(\mathbf{w})))\} d\mathbf{w}. \tag{8}$$

As in the EM and variational Bayesian algorithm for the usual mixture model, we introduce another latent (hidden) variable z_k that is 1 if the datum $\mathbf{x}$ is generated from the kth component and 0 otherwise. Then we have the following joint probability distribution of the observed data $\mathbf{x}$ and hidden variable $\mathbf{z} = (z_1, z_2, \ldots, z_K)$,

$$p(\mathbf{x}, \mathbf{z}|\omega) = \prod_{k=1}^K \{a_k \, p(\mathbf{x}|\tilde{\theta}(\mathbf{w}_k))\}^{z_k}. \tag{9}$$

The next section derives the variational Bayesian algorithm for this model.

4 Variational Bayes Method

Given n training data $X^n = \{\mathbf{x}^{(i)}\}_{i=1}^n$, using the corresponding hidden variables $Z^n = \{\mathbf{z}^{(i)}\}_{i=1}^n$, the variational Bayesian estimation approximates the Bayesian posterior distribution by the variational posterior distribution that factorizes as,

$q(Z^n, \omega) = q_1(Z^n)q_2(\omega)$. The variational posterior distribution is chosen to minimize the objective functional,

$$\tilde{F}[q] = \sum_{Z^n} \int q(Z^n, \omega) \log \frac{q(Z^n, \omega)}{p(\omega) \prod_{i=1}^n p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}|\omega)} d\omega \tag{10}$$

called the *variational free energy* (Attias 1999). The variational free energy is minimized by alternately optimizing one of $q_2(\omega)$ and $q_1(Z^n)$ while the other is fixed. The following sections give the optimal form of each distribution when the other is fixed. The derivations are omitted here (Watanabe et al. 2009).

4.1 Optimal $q_2(\omega)$ for Fixed $q_1(Z^n)$

We define

$$n_k = \textstyle\sum_{i=1}^n \langle z_k^{(i)} \rangle_{q_1(Z^n)}, \quad \nu_k = \frac{1}{n_k} \sum_{i=1}^n \langle z_k^{(i)} \rangle_{q_1(Z^n)} \mathbf{F}(\mathbf{x}^{(i)}), \tag{11}$$

where n_k is the number of data that are estimated to be generated from the kth component and ν_{kj} is the average of F_j for them. The optimal $q_2(\omega)$ for the given $q_1(Z^n)$ is $q_2(\omega) = q_2(\mathbf{a}) \prod_{k=1}^K q_2(\mathbf{w}_k)$, where

$$q_2(\mathbf{a}) = \frac{\Gamma(\sum_{k=1}^K \phi_k)}{\prod_{k=1}^K \Gamma(\phi_k)} \prod_{k=1}^K a_k^{\phi_k - 1}, q_2(\mathbf{w}_k) = \exp\{\xi_k(\mathbf{w}_k \cdot \alpha_k - G(\tilde{\theta}(\mathbf{w}_k))) - \Phi(\alpha_k, \xi_k)\}. \tag{12}$$

Here we have put $\phi_k = n_k + \phi_0$, $\alpha_k = \frac{n_k \mathrm{U}^T \nu_k + \xi_0 \alpha_0}{n_k + \xi_0}$ and $\xi_k = n_k + \xi_0$.

4.2 Optimal $q_1(Z^n)$ for Fixed $q_2(\omega)$

The optimal $q_1(Z^n)$ for the fixed $q_2(\omega)$ is given by $q_1(Z^n) \propto \prod_{i=1}^n \prod_{k=1}^K \exp(z_k^{(i)} s_k^{(i)})$, where we have defined

$$s_k^{(i)} = \Psi(\phi_k) - \Psi(\textstyle\sum_{k=1}^K \phi_k) + \frac{\partial \Phi(\alpha_k, \xi_k)}{\partial_{,k}} + \left(\frac{1}{\xi_k} \frac{\partial \Phi(\alpha_k, \xi_k)}{\partial \alpha_k}\right) \cdot \left(\mathbf{U}^T \mathbf{F}(\mathbf{x}^{(i)}) - \alpha_k\right), \tag{13}$$

and $\Psi(x) = (\log \Gamma(x))'$. Its mean is given by $\langle z_k^{(i)} \rangle_{q_1(Z^n)} = q_1(z_k^{(i)} = 1) = \frac{e^{s_k^{(i)}}}{\sum_{k=1}^K e^{s_k^{(i)}}}$. Note that the algorithm requires computing $\frac{\partial \Phi}{\partial \alpha}$ and $\frac{\partial \Phi}{\partial \xi}$ in (13).

4.3 Laplace Approximation

When the rank of the matrix $\mathbf{U}$ is L ($< M$) in general, we no longer have explicit forms of the function Φ, $\frac{\partial \Phi}{\partial \alpha}$ and $\frac{\partial \Phi}{\partial \xi}$. We provide the following approximation of Φ based on Laplace's method (Watanabe et al. 2009),

$$\Phi(\alpha, \xi) \simeq \xi(\alpha \cdot \hat{\mathbf{w}} - G(\mathbf{U}\hat{\mathbf{w}} + \mathbf{u}_0)) - \frac{L}{2} \log \frac{\xi}{2\pi} - \frac{1}{2} \log |\tilde{\mathbf{J}}|.$$

Here $\tilde{\mathbf{J}} = \mathbf{U}^T \frac{\partial^2 G(\tilde{\theta}(\mathbf{w}))}{\partial \theta \partial \theta^T} \mathbf{U}$ and $\hat{\mathbf{w}}$ is given by solving the equation $\mathbf{U}^T \frac{\partial G(\tilde{\theta}(\mathbf{w}))}{\partial \theta} = \alpha$ with respect to $\mathbf{w}$. This can be solved by an iterative method that initializes $\mathbf{w}$ with $\tilde{\mathbf{w}}$ and repeats updating $\tilde{\mathbf{w}}$ by adding

$$d\tilde{\mathbf{w}} = \tilde{\mathbf{J}}^{-1} \left(\alpha - \mathbf{U}^T \boldsymbol{\eta} \left(\mathbf{U}\tilde{\mathbf{w}} + \mathbf{u}_0 \right) \right). \tag{14}$$

The derivation is omitted here (Akaho 2004). The partial derivatives of the above expression provide approximations of $\frac{\partial \Phi}{\partial \alpha}$ and $\frac{\partial \Phi}{\partial \xi}$.

5 Dimensionality Reduction

In the previous sections, we derived the variational Bayesian learning algorithm for the mixture model whose component parameter vectors are constrained to an L-dimensional subspace. Estimating the basis vectors $\mathbf{U}$ and the displacement vector $\mathbf{u}_0$ as well enables dimensionality reduction and performing clustering simultaneously. More specifically, we propose to apply the e-PCA to cluster centers $\{\nu_k\}$ to take into account the cluster structure of the data set. In this case, the steepest descent method for minimizing $\sum_{k=1}^{K} n_k K(\theta(\nu_k) || \tilde{\theta}(\langle \mathbf{w}_k \rangle_{q_2(\mathbf{w}_k)}))$ updates the basis vector $\mathbf{u}_l$ to $\tilde{\mathbf{u}}_l$ by

$$\tilde{\mathbf{u}}_l = \mathbf{u}_l + \epsilon_u \sum_{k=1}^{K} n_k (\nu_k - \eta(\tilde{\theta}(\langle \mathbf{w}_k \rangle_{q_2(\mathbf{w}_k)}))) \langle w_{kl} \rangle_{q_2(\mathbf{w}_k)}, \tag{15}$$

where $\{n_k\}$ and $\{\nu_k\}$ are defined in (11) and ϵ_u is a small constant. We have put $w_{k0} = 1$.

We can perform dimensionality reduction and clustering simultaneously by updating the basis vectors with the above rule after once updating $q_1(Z^n)$ and $q_2(\omega)$ with the variational Bayes method described in Sect. 4. Once the subspace is estimated, the low dimensional representation of each datum can be obtained by the following expression (Watanabe et al. 2009), which provides the projection of the data point onto the subspace, $\langle \mathbf{w}^{(i)} \rangle_{p(\mathbf{w}^{(i)}|\theta^{(i)})} = \frac{1}{1+\xi_0} \frac{\partial \Phi(\alpha^{(i)}, 1+\xi_0)}{\partial \alpha}$, where $\alpha^{(i)} = \frac{\mathbf{U}^T \mathbf{F}(\mathbf{x}^{(i)}) + \xi_0 \alpha_0}{1+\xi_0}$.

6 Experiments

To demonstrate the practical applicability of the derived method, we applied it to a task of recognition of handwritten digits using the MNIST data set (LeCun et al. 1998). We used the 196-dimensional improved directional element feature vector consisting of non-negative integers (Omachi et al. 2007). This motivates the use of the mixtures of Poissons to model the probability distributions of 10 classes of digits. The n data of the training set were used for learning the mixture model for each class. We fixed a common dimensionality L over all 10 classes. We evaluated the recognition rate for the 10,000 test data by assigning each test datum to the class with the highest likelihood (posterior probability).

We first applied the full dimensional ($L = 196$) Poisson mixture model with $K = 5$ components. The component distribution is expressed in the form of the exponential family with $\mathbf{x} \in \{0, 1, \ldots\}^M$, $\mathbf{F}(\mathbf{x}) = \mathbf{x}$, $F_0(\mathbf{x}) = -\sum_{j=1}^{M} \log x_j!$ and $G(\theta) = \sum_{j=1}^{M} e^{\theta_j}$. The hyperparameters were set to $\phi_0 = 1$, $\xi_0 = 1$ and $\alpha_0 = 10 \cdot \mathbf{1}$. The results of the recognition rate are presented in Fig. 1 (Left) for different n. The results for $n = 100$ and 200 are comparable with those of the Gaussian mixture-based approach reported in Omachi et al. (2007) that is highly tuned for the discriminative task. We next applied the constrained Poisson mixture model reducing the dimensionality to $L = 2$. The hyperparameters were set in the same way as the full dimensional case. The results are presented in Fig. 1 (Left).

We can see the improvements in the recognition accuracy by reducing the dimensionality especially when the sample size is small. This demonstrates the effectiveness of the dimensionality reduction for small amounts of data. Figure 1 (Middle and Right) show the average variational free energy, (10), over the 10 classes and the error rate for different dimensionalities, $L = 2, 4, 6, 12, 18, 24, 196$, when the sample size $n = 200$ and $n = 800$ respectively. The results are averaged over five draws of data sets. The plots of error rates imply that the smaller the number n of samples, the more one needs to reduce the dimensionality L to obtain high prediction accuracy. The plots of the variational free energy show similar trends as

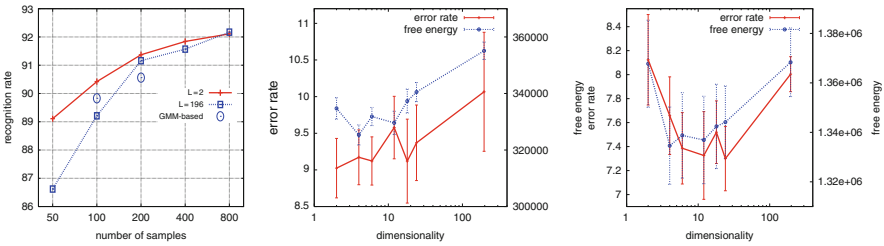


Fig. 1 *Left*: Recognition rates (%) by the Poisson mixtures with $L = 196$ (*dotted line*) and $L = 2$ (*solid line*) and those by the Gaussian mixture based method (*circle*) for different numbers of training samples. *Middle, Right*: Average variational free energy (*dotted line*) and error rates (%) (*solid line*) for different dimensionalities when $n = 200$ (*Middle*) and when $n = 800$ (*Right*)

those of the error rates implying that it provides a criterion for choosing the effective dimensionality.

7 Discussion and Conclusion

Collins et al. (2002) proposed a generalization of the PCA for exponential families. It can be viewed as a special case of the e-PCA by mapping each datum to the sample of the parameter as in (5). We gave the probabilistic density model for the e-PCA in Sect. 3 and the constrained mixture model in (6) which can be thought as a 'soft' version of the e-PCA. The semi-parametric exponential family PCA developed in Sajama and Orlitsky (2004) is the ML estimation for the constrained mixture model. The VB approach developed in this paper naturally implements complexity control by pruning redundant components and provides a criterion for choosing the effective dimensionality as presented in Sect. 6. The mixture of PCA, also known as the mixture of factor analyzers, is based on the probabilistic PCA and performs local dimensionality reduction (Tipping and Bishop 1999). In this method, multiple local latent spaces are obtained corresponding to the mixture components. The proposed method in the present paper estimates one common latent space over the components, which is more suitable for visualizing data that are multi-modally distributed in the latent space. For simultaneous dimensionality reduction and clustering, a method combining the linear discriminant analysis (LDA) with K-means has been proposed (Ding and Li 2007). Compared to such methods, a key feature of the algorithm derived in this paper is to use the distance between probability distributions as the measure of similarity in the space of data. Extending the LDA to incorporate the exponential family and introducing it to the estimation of the basis vectors would be an important undertaking in the future.

References

Akaho, S. (2004). e-PCA and m-PCA: Dimension reduction of parameters by information geometry. *Proceedings of IJCNN*, 129–134.

Amari, S., & Nagaoka, H. (2000). *Methods of Information Geometry*. Oxford: AMS and Oxford University Press.

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. *Proceedings of UAI*, 21–30.

Banerjee, A., Merugu, S., Dhillon, I. S., & Ghosh J. (2005). Clustering with Bregman Divergences. *Journal of Machine Learning Research, 6*, 1705–1749.

Bishop, C. M. (1999). Bayesian PCA. *Advances in NIPS,11*, 382–388.

Collins, M., Dasgupta, S., & Schapire R. (2002). A generalization of principal component analysis to the exponential family. *Advances in NIPS,14*, 617–624.

Ding, C., & Li, T. (2007). Adaptive dimension reduction using discriminant analysis and K-means clustering. *Proceedings of ICML*, 521–528.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner P. (1998). Gradient-based learning applied to document recognition. *Proceedongs of the IEEE, 86*(11), 2278–2324.

Omachi, S., Omachi, M., & Aso H. (2007). An approximation method of the quadratic discriminant function and its application to estimation of high-dimensional distribution. *IEICE Transitions on Information System, E90-D*(8), 1160–1167.

Sajama, & Orlitsky, A. (2004). Semi-parametric exponential family PCA : Reducing dimensions via non-parametric latent distribution estimation. *Technical Report CS2004-0790*, University of California at San Diego.

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B, 61*, 611–622.

Watanabe, K., Akaho, S., Omachi, S., & Okada, M. (2009). Variational bayesian mixture model on a subspace of exponential family distributions. *IEEE Trans. Neural Networks, 20*(11), 1783–1796.

A Partitioning Method for the Clustering of Categorical Variables

Marie Chavent, Vanessa Kuentz, and Jérôme Saracco

Abstract In the framework of clustering, the usual aim is to cluster observations and not variables. However the issue of clustering variables clearly appears for dimension reduction, selection of variables or in some case studies. A simple approach for the clustering of variables could be to construct a dissimilarity matrix between the variables and to apply classical clustering methods. But specific methods have been developed for the clustering of variables. In this context center-based clustering algorithms have been proposed for the clustering of quantitative variables. In this article we extend this approach to categorical variables. The homogeneity criterion of a cluster of categorical variables is based on correlation ratios and Multiple Correspondence Analysis is used to determine the latent variable of each cluster. A simulation study shows that the method recovers well the underlying simulated clusters of variables. Finally an application on a real data set also highlights the practical benefits of the proposed approach.

1 Introduction

From a general point of view, variable clustering lumps together variables which are strongly related to each other and thus bring the same information. It is a possible solution for selection of variables or dimension reduction which are current problems with the emergency of larger and larger data bases. In some case studies, the main objective is to cluster variables and not units, such as sensory analysis (identification of groups of descriptors), biochemistry (gene clustering), etc. Techniques of variable clustering can also be useful for association rules mining (see for instance Plasse et al. 2007).

V. Kuentz (✉)
Université de Bordeaux, IMB, CNRS, UMR 5251, France
and
INRIA Bordeaux Sud-Ouest, CQFD team, France
e-mail: kuentz@math.u-bordeaux1.fr

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_9,

A simple approach for the clustering of variables could be to calculate first the matrix of the dissimilarities between the variables and then to apply classical clustering methods which are able to deal with dissimilarity matrices (complete or average linkage hierarchical clustering among others). Other methods like Ward or k-means (dealing only with quantitative data) could also be applied on the numerical coordinates obtained from Multidimensional Scaling of this dissimilarity matrix. But specific methods have also been developed for the clustering of variables. In this context Cluster Analysis of Variables Around Latent Components (Vigneau and Qannari 2003) and Diametrical clustering (Dhillon et al. 2003) are two independently proposed center-based clustering methods for the clustering of quantitative variables. These methods are iterative two steps relocation algorithms involving at each iteration the identification of a cluster centroid by optimization of an homogeneity criterion and the allocation of each variable to the "nearest" cluster. The cluster centroid is a synthetic component, called latent variable, which summarizes the variables belonging to the cluster. When high absolute correlations imply agreement, both methods aim at maximizing the same homogeneity criterion (based on squared correlations). In this case, the latent variable of a cluster is the first principal component issued from Principal Component Analysis (PCA) of the matrix containing the variables of the cluster.

In this paper we extend this relocation partitioning method to the case of categorical variables. The homogeneity criterion is now based on correlation ratios between the categorical variables and the cluster centroids which are numerical variables, defined by optimization of this homogeneity criterion.

Section 2 presents the center-based clustering algorithm for the clustering of categorical variables. A simulation study is carried out in Sect. 3 to show the numerical performance of the approach and a real data application illustrates its practical benefits. Finally some concluding remarks are given in Sect. 4.

2 A Center-Based Partitioning Method for the Clustering of Categorical Variables

Let $\mathbf{X} = (x_{ij})$ be a data matrix of dimension (n, p) where a set of n objects are described on a set of p categorical variables, that is, $x_{ij} \in \mathcal{M}_j$ where $\mathcal{M}_j$ is the set of categories of the jth variable. Let $\mathcal{V} = \{\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_p\}$ be the set of the p columns of $\mathbf{X}$, called for sake of simplicity categorical variables. We denote by $\mathcal{P} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k, \ldots, \mathcal{C}_K\}$ a partition of $\mathcal{V}$ into K clusters and by $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_k, \ldots, \mathbf{y}_K\}$ a set of K vectors of $\mathbb{R}^n$ called latent variables.

The aim is to find a couple $(\mathcal{P}, \mathcal{Y})$, optimum with respect to the following homogeneity criterion:

$$H(\mathcal{P}, \mathcal{Y}) = \sum_{k=1}^{K} S(\mathcal{C}_k, \mathbf{y}_k), \tag{1}$$

where S measures the adequacy between $\mathcal{C}_k$ and the latent variable $\mathbf{y}_k$:

$$S(\mathcal{C}_k, \mathbf{y}_k) = \sum_{\mathbf{x}_j \in \mathcal{C}_k} \eta^2(\mathbf{x}_j, \mathbf{y}_k), \tag{2}$$

with $\eta^2(\mathbf{x}_j, \mathbf{y}_k)$ the correlation ratio measuring the link between $\mathbf{x}_j$ and $\mathbf{y}_k$.

Definition 1. The correlation ratio $\eta^2(\mathbf{x}_j, \mathbf{y}_k) \in [0, 1]$ is equal to the between group sum of squares of $\mathbf{y}_k$ in the groups defined by the categories of $\mathbf{x}_j$, divided by the total sum of squares of $\mathbf{y}_k$. We have with $\mathbf{y}_k = (y_{k,1}, \ldots, y_{k,i}, \ldots, y_{k,n}) \in \mathbb{R}^n$, $\eta^2(\mathbf{x}_j, \mathbf{y}_k) = \dfrac{\sum_{s \in \mathcal{M}_j} n_s (\bar{\mathbf{y}}_{ks} - \bar{\mathbf{y}}_k)^2}{\sum_{i=1}^{n} (y_{k,i} - \bar{\mathbf{y}}_k)^2}$, with n_s the frequency of category s, $\mathcal{M}_j$ the set of categories of $\mathbf{x}_j$ and $\bar{\mathbf{y}}_{ks}$ the mean value of $\mathbf{y}_k$ calculated on the objects belonging to category s.

2.1 Definition of the Latent Variable

The latent variable $\mathbf{y}_k$ of a cluster $\mathcal{C}_k$ is defined by maximization of the adequacy criterion S:

$$\mathbf{y}_k = \arg\max_{\mathbf{u} \in \mathbb{R}^n} \sum_{\mathbf{x}_j \in \mathcal{C}_k} \eta^2(\mathbf{x}_j, \mathbf{u}). \tag{3}$$

Proposition 1. *The first principal component obtained with Multiple Correspondence Analysis of $\mathbf{X}_k$, the matrix containing the variables of $\mathcal{C}_k$, is a solution of* (3) *and is then a latent variable $\mathbf{y}_k$ of $\mathcal{C}_k$.*

Proof. Let us introduce some notations. Let $\mathbf{G} = (g_{is})_{n \times q_k}$, with $g_{is} = 1$ if i belongs to category s and 0 otherwise, be the indicator matrix of the q_k categories of the p_k variables in $\mathcal{C}_k$. We note $\mathbf{F}_k = (f_{is})_{n \times q_k}$ the frequency matrix built from $\mathbf{G}$. The row and column marginals define respectively the vectors of row and column masses $\mathbf{r}_k$ and $\mathbf{c}_k$. The ith element of $\mathbf{r}_k$ is $f_{i.} = \frac{1}{n}$ and the sth element of $\mathbf{c}_k$ is $f_{.s} = \frac{n_s}{np_k}$. Let us consider the two following diagonal matrices $\mathbf{D}_n = \mathrm{diag}(\mathbf{r}_k)$ and $\mathbf{D}_{q_k} = \mathrm{diag}(\mathbf{c}_k)$. We introduce the matrix $\widetilde{\mathbf{F}}_k = \mathbf{D}_n^{-1/2}(\mathbf{F}_k - \mathbf{r}_k \mathbf{c}_k^t)\mathbf{D}_{q_k}^{-1/2}$ which general term writes:

$$\tilde{f}_{is} = \frac{\sqrt{n_s p_k}}{n_s}\left(\frac{g_{is}}{p_k} - \frac{n_s}{np_k}\right) = \begin{cases} \frac{\sqrt{n_s p_k}}{n_s}\left(\frac{1}{p_k} - \frac{n_s}{np_k}\right) & \text{if } i \text{ belongs to category } s, \\ 0 & \text{otherwise.} \end{cases}$$

First we show that if $\mathbf{u}^t\mathbf{u} = 1$ and $\bar{\mathbf{u}} = 0$, then $\frac{1}{p_k}\sum_{\mathbf{x}_j \in \mathcal{C}_k} \eta^2(\mathbf{x}_j, \mathbf{u}) = \mathbf{u}^t \widetilde{\mathbf{F}}_k \widetilde{\mathbf{F}}_k^t \mathbf{u}$. If $\bar{\mathbf{u}} = 0$, $\sum_{i=1}^{n} \tilde{f}_{is}\mathbf{u}_i = \frac{\sqrt{n_s}}{\sqrt{p_k}}\bar{\mathbf{u}}_s$, where $\bar{\mathbf{u}}_s$ is the mean value of $\mathbf{u}$ calculated on the objects belonging to category s. Thus we have:

$$\mathbf{u}^t\widetilde{\mathbf{F}}_k\widetilde{\mathbf{F}}_k^t\mathbf{u} = \frac{1}{p_k}\sum_{\mathbf{x}_j\in\mathcal{C}_k}\sum_{s\in\mathcal{M}_j} n_s\bar{\mathbf{u}}_s^2 = \frac{\frac{1}{p_k}\sum_{\mathbf{x}_j\in\mathcal{C}_k}\sum_{s\in\mathcal{M}_j}\frac{n_s}{n}(\bar{\mathbf{u}}_s-0)^2}{\frac{1}{n}}$$
$$= \frac{1}{p_k}\sum_{\mathbf{x}_j\in\mathcal{C}_k}\eta^2(\mathbf{x}_j,\mathbf{u}).$$

As the first normalized eigenvector $\mathbf{u}_1$ of $\widetilde{\mathbf{F}}_k\widetilde{\mathbf{F}}_k^t$ maximizes $\mathbf{u}^t\widetilde{\mathbf{F}}_k\widetilde{\mathbf{F}}_k^t\mathbf{u}$, it is a solution of (3).

Finally, as $\eta^2(\mathbf{x}_j,\mathbf{u}) = \eta^2(\mathbf{x}_j,\alpha\mathbf{u})$, for any nonnull real α, $\alpha\mathbf{u}_1$ is also a solution of (3). The proof is then completed by showing that $\mathbf{u}_1$ is colinear to the first principal component issued from MCA on the centered row profiles matrix $\mathbf{R}_k$ of $\mathbf{X}_k$. MCA can be viewed as a weighted PCA applied to $\mathbf{R}_k = \mathbf{D}_n^{-1}(\mathbf{F}_k - \mathbf{r}_k\mathbf{c}_k^t)$. The first principal component is then $\boldsymbol{\psi}_1 = \mathbf{R}_k\mathbf{D}_{q_k}^{-1/2}\mathbf{v}_1$, where $\mathbf{v}_1$ is the eigenvector associated with the largest eigenvalue λ_1 of $\widetilde{\mathbf{F}}_k^t\widetilde{\mathbf{F}}_k$. Then we use the SVD of $\widetilde{\mathbf{F}}_k$ to write $\boldsymbol{\psi}_1 = \sqrt{\lambda_1}\sqrt{n}\mathbf{u}_1$, and the proof is complete. □

2.2 The Center-Based Clustering Algorithm

The corresponding center-based algorithm is the following:

(a) *Initialization step*: We compute the first K principal components issued from MCA of $\mathbf{X}$. Then we assign each variable to the nearest component, that is to the component with which its correlation ratio is the highest. Thus we get an initial partition $\{C_1,\ldots,C_k,\ldots,C_K\}$ of $\mathcal{V}$.

(b) *Representation step*: $\forall k = 1,...,K$, compute the latent variable $\mathbf{y}_k$ of $\mathcal{C}_k$ as the first principal component $\boldsymbol{\psi}_1$ of $\mathbf{X}_k$ (or as the first normalized eigenvector $\mathbf{u}_1$ of $\widetilde{\mathbf{F}}_k\widetilde{\mathbf{F}}_k^t$).

(c) *Allocation step*: $\forall j = 1,...,p$, find ℓ such that $\ell = \arg\max_{k=1,\ldots,K}\eta^2(\mathbf{x}_j,\mathbf{y}_k)$. Let $\mathcal{C}_k$ be the previous cluster of $\mathbf{x}_j$. Then if $\ell \neq k$, $\mathcal{C}_\ell \leftarrow \mathcal{C}_\ell \cup \{\mathbf{x}_j\}$ and $\mathcal{C}_k \leftarrow \mathcal{C}_k\backslash\{\mathbf{x}_j\}$.

(d) If nothing changes in (c) then *stop*, else return to step (b).

Proposition 2. *The center-based algorithm converges to a local optimum of the homogeneity criterion H.*

Proof. We show that the homogeneity criterion H increases until convergence. For that we have to prove that $H(\mathcal{P}^n,\mathcal{Y}^n) \leq H(\mathcal{P}^n,\mathcal{Y}^{n+1}) \leq H(\mathcal{P}^{n+1},\mathcal{Y}^{n+1})$, where the superscript n denotes the nth iteration of the algorithm.

The first inequality is verified since the latent variable of a cluster $\mathcal{C}_k^n$ is defined to maximize S and then $S(\mathcal{C}_k^n,\mathbf{y}_k^n) \leq S(\mathcal{C}_k^n,\mathbf{y}_k^{n+1})$. Then by summing up on k, we get $H(\mathcal{P}^n,\mathcal{Y}^n) \leq H(\mathcal{P}^n,\mathcal{Y}^{n+1})$.

Finally according to the definition of the allocation step, we have $\sum_{k=1}^{K} \sum_{\mathbf{x}_j \in C_k^n} \eta^2(\mathbf{x}_j, \mathbf{y}_k^{n+1}) \leq \sum_{k=1}^{K} \sum_{\mathbf{x}_j \in C_k^{n+1}} \eta^2(\mathbf{x}_j, \mathbf{y}_k^{n+1})$, which proves the second inequality. □

3 Applications

In this section we present some applications of the center-based clustering algorithm for the clustering of categorical variables. In the first one we consider a simulated example in order to show the numerical performance of the proposed approach. Then we apply it on a real categorical data set to show the potential of the approach.

3.1 Simulation Study

In this simulation study we consider six binary variables $x_1, \ldots, x_6$ and we study four different states of relationship between them. The idea is to simulate at first three groups of variables which are well defined, that is the variables within each cluster are strongly linked to each other and they are weakly related to variables belonging to other clusters. They form the partition $\mathcal{Q} = (\mathcal{Q}_1, \mathcal{Q}_2, \mathcal{Q}_3)$ with $\mathcal{Q}_1 = \{x_1, x_2\}$, $\mathcal{Q}_2 = \{x_3, x_4\}$ and $\mathcal{Q}_3 = \{x_5, x_6\}$. Then we increasingly disrupt the underlying structure. Let a (resp. b, c, d, e, f) denote a category of x_1 (resp. x_2, x_3, x_4, x_5, x_6) and $\mathbb{P}$ denote a probability measure. To generate a contingency table, the following log-linear model (see for instance Agresti 2002) is simulated:

$$\begin{aligned}\log(\mathbb{P}(x_1 = a, \ldots, x_6 = f)) = (\lambda_a^{x_1} + \lambda_b^{x_2} + \beta_{ab}^{x_1 x_2}) + (\lambda_c^{x_3} + \lambda_d^{x_4} + \beta_{cd}^{x_3 x_4}) \\ + (\lambda_e^{x_5} + \lambda_f^{x_6} + \beta_{ef}^{x_5 x_6}) + \beta_{ad}^{x_1 x_4} + \beta_{cf}^{x_3 x_6} \quad (4)\end{aligned}$$

where $a, b, c, d, e, f \in \{0, 1\}$. The parameters $\lambda_a^{x_1}, \lambda_b^{x_2}, \lambda_c^{x_3}, \lambda_d^{x_4}, \lambda_e^{x_5}, \lambda_f^{x_6}$ represent the effect of each variable and the parameters $\beta_{ab}^{x_1 x_2}, \beta_{cd}^{x_3 x_4}, \beta_{ef}^{x_5 x_6}$ are interactions corresponding with cohesion terms in each group. The parameter $\beta_{ad}^{x_1 x_4}$ (resp. $\beta_{cf}^{x_3 x_6}$) is used to add some interactions between categories of variables belonging to different groups $\mathcal{Q}_1$ and $\mathcal{Q}_2$ (resp. $\mathcal{Q}_2$ and $\mathcal{Q}_3$). The first state of mixing corresponds to the initial partition and is called "no mixing". Then we moderately mix the two groups by increasing the value of $\beta_{00}^{x_1 x_4}$, it will be referred as "moderate mixing". In the third case named "strong mixing", the value of $\beta_{00}^{x_1 x_4}$ is high. In the last state called "very strong mixing", the values of $\beta_{00}^{x_1 x_4}$ and $\beta_{00}^{x_3 x_6}$ are high. Thus there is no more structure in the data.

For each state of mixing we simulate $N = 50$ contingency tables, each corresponding to a global sample size $n = 2{,}000$ using log-linear model (4), where the values of the parameters are given in Table 1. Only the nonnull parameter values

Table 1 Values of the parameters of model (4) used in the simulations

State of mixing	No mixing	Moderate mixing	Strong mixing	Very strong
Effect of each variable	$\lambda_0^{x_1} = \lambda_0^{x_3} = \lambda_0^{x_5} = 1$ $\lambda_0^{x_2} = \lambda_0^{x_4} = \lambda_0^{x_6} = h \in [1, 1.5]$			
Cohesion terms	$\beta_{00}^{x_1x_2} = -1.5$ $\beta_{00}^{x_3x_4} = -1.1$ $\beta_{00}^{x_5x_6} = -0.9$	$\beta_{00}^{x_1x_2} = -1.5$ $\beta_{00}^{x_3x_4} = -1.2$ $\beta_{00}^{x_5x_6} = -1$	$\beta_{00}^{x_1x_2} = -0.8$ $\beta_{00}^{x_3x_4} = -0.7$ $\beta_{00}^{x_5x_6} = -0.9$	
Interaction terms	0	$\beta_{00}^{x_1x_4} = -0.9$	$\beta_{00}^{x_1x_4} = -1.5$	$\beta_{00}^{x_1x_4} = 0.9$ $\beta_{00}^{x_3x_6} = -1.5$

are specified, all the remaining ones are set to zero. In this table the value h of the effect parameters $\lambda_0^{x_2}, \lambda_0^{x_4}, \lambda_0^{x_6}$ is generated with the univariate uniform distribution on $[1, 1.5]$ to get N slightly different contingency tables.

We apply the proposed algorithm on the generated categorical data.

- When there is no mixing between the groups, the proposed approach always recovers the underlying clusters.
- When the mixing between the groups is moderate, the algorithm misclassifies one variable. We always obtain the partition $\{\{x_1, x_2, x_4\}, \{x_3\}, \{x_5, x_6\}\}$.
- When two groups are strongly mixed, the algorithm always misclassifies two variables. The corresponding partition is $\{\{x_1, x_4\}, \{x_2, x_3\}, \{x_5, x_6\}\}$.
- When the mixing is very strong, not surprisingly the algorithm misclassifies three variables since there is no more visible structure in the data. The obtained partition is always $\{\{x_1\}, \{x_2, x_4, x_5\}, \{x_3, x_6\}\}$.

3.2 Real Data Application

We consider a real data set on a user satisfaction survey of pleasure craft operators on the "Canal des Deux Mers" located in South of France which contains numerous questions with numerical or categorical answers. This study has been realized from June to September 2008. In this application we only focus on 14 categorical variables described in Table 2. The sample size is $n = 709$ pleasure craft operators.

In this case study, we have chosen to retain $K = 5$ clusters because it provides a satisfactory mean correlation ratio value (0.68), that is the mean of the correlation ratio between the variables in each cluster and the corresponding latent variable. Moreover the interpretation of the clusters seems to be sound. This choice has also been confirmed by a bootstrap approach which consists in generating multiple data replications of the data set and examining if the partition is stable. Table 3 describes the five-clusters partition of the variables. For instance cluster 4 contains variables dealing with the use of the canal. As has already been pointed, MCA is used to have a first solution to start the algorithm. Comparing the obtained solution with the MCA solution shows that cluster 1 and 4 are merged and that only one iteration is needed to obtain convergence to a local optimum corresponding to the partition

Table 2 Description of the 14 categorical variables

Name of the variable	Description of the variable	Categories
x_1 = "sites worth visiting"	*What do you think about information you were provided with concerning sites worth visiting?*	Satisfactory, unsatisfactory, no opinion
x_2 = "leisure activity"	*How would you rate the information given on leisure activity?*	
x_3 = "historical canal sites"	*What is your opinion concerning tourist information on historical canal sites (locks, bridges, etc.)?*	
x_4 = "manoeuvres"	*At the start of your cruise, were you sufficiently aware of manoeuvres at locks?*	
x_5 ="authorized mooring"	*At the start of your cruise, were you sufficiently aware of authorized mooring?*	Yes, no
x_6 = "safety regulations"	*At the start of your cruise, were you sufficiently aware of safety regulations?*	
x_7 = "services"	*Please give us your opinion about signs you encountered along the way concerning information regarding services.*	Satisfactory, unsatisfactory
x_8 = "number of taps"	*What do you think about number of taps on your trip?*	Sufficient, unsufficient
x_9 = "cost of water"	*The general cost of water is ...*	Inexpensive, average, expensive
x_{10} = "cost of electricity"	*The general cost of electricity is ...*	
x_{11} = "visibility of electrical outlets"	*What is your opinion of visibility of electrical outlets?*	Sufficient, unsufficient
x_{12} = "number of electrical outlets"	*What do you think about number of electrical outlets on your trip?*	
x_{13} = "cleanliness"	*How would you describe the canal's degree of cleanliness?*	Clean, average, dirty
x_{14} = "unpleasant odours"	*Were there unpleasant odours on the canal?*	None, occasional, frequent

Table 3 Partition of the 14 categorical variables into five clusters (correlation ratio between the variable and the latent variable of the cluster)

$\mathcal{C}_1$: **environment**	$\mathcal{C}_2$: **navigation rules**	$\mathcal{C}_3$: **cost of services**
Cleanliness (0.68)	Manoeuvres (0.66)	Cost of water (0.84)
Unpleasant odours (0.68)	Authorized mooring (0.71)	Cost of electricity (0.84)
	Safety regulations (0.69)	
$\mathcal{C}_4$: use of the canal	$\mathcal{C}_5$: available services	
Sites worth visiting (0.71)	Services (0.40)	
leisure activity (0.69)	Number of taps (0.59)	
historical canal sites (0.46)	Visibility of electrical outlets (0.65)	
	Number of electrical outlets (0.71)	

Table 4 Values of the Tschuprow coefficient between the variables of cluster 4 and the remaining ones

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	...	x_{14}
x_1	**1.00**	**0.36**	**0.24**	0.09	0.10	0.11	0.08	0.06	...	0.05
x_2	**0.36**	**1.00**	**0.20**	0.10	0.11	0.13	0.11	0.07	...	0.03
x_3	**0.24**	**0.20**	**1.00**	0.02	0.04	0.05	0.11	0.08	...	0.05

given in Table 3. The value in brackets of this table corresponds to the correlation ratio between the variable and the latent variable representing the cluster it belongs to. We see that the variables in a cluster are highly related with their latent variable. Table 4 gives the values of the Tschuprow coefficient between the variables of cluster 4 $\{x_1, x_2, x_3\}$ and the remaining ones. We see that the variables are more related with variables in the same cluster than with the variables in the other clusters. This means that dimension reduction is possible. For instance in this case study we could reduce the number of the questions in the survey by selecting one question in each cluster. Furthermore we could replace the classical previous step of MCA for the clustering of the individuals by the construction of the latent variables.

4 Concluding Remarks

In this paper we propose an extension of an existing center-based algorithm to the case of categorical variables. For numerical variables the homogeneity criterion is calculated with squared correlations between the variables of the cluster and its latent variable, which is defined as the first principal component issued from PCA. For categorical variables correlation ratios and MCA are then used respectively in place of squared correlations and PCA. The originality of the proposed approach lies in the fact that the center of a cluster of categorical variables is a numerical variable. A simulation study shows that the proposed method is efficient to recover simulated clusters of variables and a real data application illustrates the practical benefits of the approach.

The initialization of the algorithm is actually reached by computing the first K principal components issued from MCA. Another solution is to run several times the algorithm with multiple random initializations and to retain the best partition in sense of the homogeneity criterion. The initialization with MCA can also be coupled with a rotation to start with a better partition. For instance, the planar iterative rotation procedure proposed for MCA by Chavent et al. (2009) can be used. Another interesting perspective would be to use this partitioning method in a divisive hierarchical approach to divide at best a cluster into two sub-clusters. Both research on ascendant and divisive hierarchical algorithms and a comparison of the different types of initialization for the partitioning method are currently under investigation.

Source codes of the implementation in `R` are available from the authors.

Acknowledgements The authors are greatful to the public corporation "Voies Navigables de France" and the private firm Enform for providing the real data set.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.) New York: Wiley Series in Probability and Statistics.

Chavent, M., Kuentz, V., & Saracco J. (2009). *Rotation in multiple correspondence analyis: a planar rotation iterative procedure*, *Submitted paper*.

Dhillon, I. S., Marcotte, E. M., & Roshan, U. (2003). Diametrical clustering for identifying anti-correlated gene clusters. *Bioinformatics, 19*(13), 1612–1619.

Plasse, M., Nianga, N., Saporta, G., Villeminot, A., & Leblond, L. (2007). Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. *Computational Statistics and Data Analysis, 52*(1), 596–613.

Vigneau, E., & Qannari, E. M. (2003). Clustering of variables around latent components. *Communications in Statistics Simulation and Computation, 32*(4), 1131–1150.

Treed Gaussian Process Models for Classification

Tamara Broderick and Robert B. Gramacy

Abstract Recognizing the success of the treed Gaussian process (TGP) model as an interpretable and thrifty model for nonstationary regression, we seek to extend the model to classification. By combining Bayesian CART and the latent variable approach to classification via Gaussian processes (GPs), we develop a Bayesian model averaging scheme to traverse the full space of classification TGPs (CTGPs). We illustrate our method on synthetic and real data and thereby show how the combined approach is highly flexible, offers tractable inference, produces rules that are easy to interpret, and performs well out of sample.

1 Introduction and Background

A Gaussian process (GP) (Rasmussen and Williams 2006) is a popular nonparametric model for regression and classification that specifies a prior over functions. For ease of computation, typical priors often confine the functions to stationarity. While stationarity is a reasonable assumption for many data sets, still many exhibit only local stationarity. A treed Gaussian process (TGP) (Gramacy and Lee 2008) represents a thrifty alternative (for the regression problem) that takes a local divide-and-conquer approach to nonstationary modeling. It defines a treed partitioning process on the predictor space and fits separate stationary GPs to the regions at the leaves. The treed form of the partition makes the model particularly interpretable.

We seek to extend the TGP model to classification. Separately, both treed models [CART (Breiman et al. 1984) and Bayesian CART (Chipman et al. 1998)] and GPs (Neal 1998) have already been successfully applied to classification. The machinery of treed partitions for a nonstationarity process and latent variables for classification suggests two possible combinations. We argue that one of the two offers clear advantages in terms of faster mixing in the resulting trans-dimensional Markov chain. Furthermore, we explore schemes for efficiently sampling the latent variables, which

T. Broderick (✉)
Statistical Laboratory, University of Cambridge, Cambridge, UK
e-mail: tb361@statslab.cam.ac.uk

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_10,

is important to obtain good mixing in the (significantly) expanded parameter space compared to the regression case. Before delving further into the details we shall review the GP model for regression and classification.

1.1 Gaussian Processes for Regression and Classification

For real-valued p-dimensional inputs, a Gaussian process (GP) is formally a prior on the space of functions $Z : \mathbb{R}^p \to \mathbb{R}$ such that the function values $Z(x)$ at any finite set of input points x have a joint Gaussian distribution Stein (1999). A particular GP is defined by its mean and correlation functions. The mean function $\mu(x) = \mathbb{E}(Z(x))$ is often constant or linear in the explanatory variable coordinates: $\mu(x) = f(x)\beta$, where $f(x) = [1, x]$. The correlation function is defined as $K(x, x') = \sigma^{-2}[Z(x) - \mu(x)]^\top [Z(x') - \mu(x')]$. We follow Gramacy and Lee (2008) and further assume that the correlation function can be decomposed into two components: a underlying strict correlation function K^* and a noise term of constant and strictly positive size g that is i.i.d. at the predictor points: $K(x_i, x_j) = K^*(x_i, x_j) + g\delta_{i,j}$. Here, $\delta_{i,j}$ is the Kronecker delta function, and g is called the *nugget*. It represents a source of measurement error and can offer improved numerical stability. A popular choice for $K^*(x, x')$ is the anisotropic squared exponential correlation:

$$K^*(x, x') = \exp\left\{-\sum_{p=1}^{P} \frac{(x_p - x'_p)^2}{d_p}\right\}.$$

The strictly positive parameters d_p describe the *range* (or *length-scale*) of the process in each direction. Further discussion of appropriate correlation structures for GPs is provided by, e.g., Stein (1999). The GP model features some notable drawbacks, including stationarity and computational cost (requiring the $O(N^3)$ inversion of an $N \times N$ matrix).

We may extend the GP model for regression to classification by introducing latent variables (Neal 1998). Here, the data consist of predictors X and classes $C \in \{1, \ldots M\}$. For each class, we define a set of latent variables $\{Z_m\}_{m=1}^M$. For a particular class m, the latent variable generative model is a GP as before: $Z_m \sim \mathcal{N}(\mu_m(X), K_m(X, X))$. The class probabilities are now obtained from the latent variables via a softmax function:

$$p(C(x) = m) \propto \exp(-Z_m(x)). \tag{1}$$

Finally, the classes are drawn from a categorical distribution with these probabilities. In practice, we eliminate redundancy by including only $M - 1$ GPs and then set the last set of latent variables to zero. Similar drawbacks to GPs apply in the classification context, with the added complexity of $O(MN)$ extra latent variables that need to be estimated. Many of these issues are addressed by partitioning.

2 Treed Gaussian Processes

Fitting different, independent models to the data in separate regions of the input space naturally implements a globally nonstationary model. Moreover, dividing up the space results in smaller local covariance matrices, which are more quickly inverted. Finally, partitions offer a natural and data-inspired blocking strategy for latent-variable sampling in classification.

2.1 TGP for Regression

Treed models provide a partition process that is recursive, so arbitrary axis-aligned regions in the p-dimensional predictor space may be defined. Conditional on a treed partition, models are fit in each of the leaf regions. In CART (Breiman et al. 1984) the underlying models are "constant" in that only the mean and standard deviation of the real-valued outputs are inferred. The tree is "grown" according to one of many decision-theoretic heuristics and may be "pruned" using cross-validation methods. In Bayesian CART (BCART), these models may be either constant (Chipman et al. 1998) or linear (Chipman et al. 2002) and, by contrast with CART, the partitioning structure is determined by Monte Carlo inference on the joint posterior of the tree and the models used at the leaves. In regression TGP (hereafter RTGP), the leaf models are GPs, but otherwise the setup is identical to BCART. Note that the constant and linear model are just special cases of the GP model. Thus RTGPs encompass BCART for regression, and inference may proceed according to a nearly identical Monte Carlo method, described shortly.

The hierarchical model for the RTGP begins with the tree prior, following Chipman et al. (1998). Let $r \in \{1, \ldots, R\}$ index the R non-overlapping regions partitioned by the tree $\mathcal{T}$ drawn from the tree-prior. In the regression problem, each region contains data $\{X_r, Z_r\}$. Let a be the number of columns and n_r the number of rows in F_r, extending the predictor matrix to include an intercept term: $F_r = (1, X_r)$. A "constant mean" may be obtained with $F_r = 1$; in this case, $a = 1$. The generative model for the GP in region r incorporates the multivariate normal ($\mathcal{N}$), inverse-gamma (IG), and Wishart (W) distributions:

$$\begin{aligned} Z_r|\beta_r, \sigma_r^2, K_r &\sim \mathcal{N}_{n_r}(F_r\beta_r, \sigma_r^2 K_r) & \sigma_r^2 &\sim IG(\alpha_\sigma/2, q_\sigma/2) \\ \beta_r|\sigma_r^2, \tau_r^2, W, \beta_0 &\sim \mathcal{N}_a(\beta_0, \sigma_r^2\tau_r^2 W) & \beta_0 &\sim \mathcal{N}_a(\mu, B) \\ \tau_r^2 &\sim IG(\alpha_\tau/2, q_\tau/2) & W^{-1} &\sim W((\rho V)^{-1}, \rho). \end{aligned} \tag{2}$$

The hyperparameters $\mu, B, V, \rho, \alpha_\sigma, q_\sigma, \alpha_\tau, q_\tau$ are constant in the model.

We sample from the joint distribution of the tree structure $\mathcal{T}$, the R sets of GP parameters θ_r ($r = 1, \ldots, R$) in each region defined by $\mathcal{T}$, and the GP hyperparameters θ_0 (those variables in (2) that are not treated as constant but also not indexed by r) by Markov Chain Monte Carlo (MCMC). We sequentially draw $\theta_0|\text{rest}$, $\theta_r|\text{rest}$ for each $r = 1, \ldots, R$, and $\mathcal{T}|\text{rest}$. Conditional on $\mathcal{T}$, all parameters

$(\theta_r, r = 1, \ldots, R)$ and hyperparameters of the GPs can be sampled with Gibbs steps, with the exception of the covariance function parameters $\{d_r, g_r\}$. Expressions are provided by Gramacy and Lee (2008).

Monte Carlo integration over tree space, conditional on the GP parameters θ_r, $r = 1, \ldots, R$, is more involved since the new tree structure drawn from the distribution $\mathcal{T}|\text{rest}$ may have a different number of leaf nodes than its predecessor. Changing the number of leaf nodes changes the dimension of $\theta = (\theta_1, \ldots, \theta_R)$, so simple MH draws are insufficient in this case. Instead, reversible jump Markov Chain Monte Carlo (RJ-MCMC) allows a principled transition between models of different sizes (Richardson and Green 1997). In this framework, we mostly use the same four moves (*grow, prune, change, swap*) as in BCART to explore the tree space. These moves and their application are described in more detail in Gramacy and Lee (2008). From the hierarchical model in (2) we can solve for the predictive distribution of the outputs Z. These are expressed in closed form by Gramacy and Lee (2008) and are provided later in this section (3) in the particular context of sampling from the latent variables used for classification.

2.2 TGP for Classification

One can envision at least two possible ways in which the latent variable approach and the treed approach to classification may be combined. The first method starts with an RTGP; recall that the RTGP tree partitions the predictor space into regions, each of which is assigned a stationary regression model (the GP). Analogously in the classification case, we could partition the predictor space into regions with a single tree and assign a stationary classification model (the CGP) to each region. Since this model contains just *one* tree, we call it the OTGP. The second method starts from the CGP; recall that the CGP uses $M - 1$ sets of real-valued latent variables generated from $M - 1$ stationary regression models (GPs). To introduce nonstationarity, instead consider $M-1$ sets of real-valued latent variables generated from $M - 1$ nonstationary regression models (RTGPs). Since this model contains *multiple* trees, we call it the MTGP.

The MTGP has a variety of advantages over the OTGP. Indeed, the OTGP is a special case of the MTGP where the parameters, hyperparameters, tree structure, and hence region partitions are fixed across all trees. Thus, the MTGP represents natural splits in the data more easily and more interpretably. The MTGP requires fewer splits per tree than the OTGP. The MTGP also enjoys a higher Monte Carlo acceptance rate in tree space since, compared to OTGP moves, its moves are more local. In OTGP the data across all classes contribute to the acceptance probability, whereas in MTGP the acceptance of moves depends on how the particular class involved may be distinguished from all others. These last two considerations combine to ensure better mixing for the MTGP. Finally, from a practical standpoint, the MTGP is more directly implemented from the RTGP as it essentially amalgamates $M - 1$ of these models. Thus, we focus on the MTGP in what follows and refer to it as the CTGP (TGP for classification) in analogy to the acronym RTGP.

The hierarchical model for the CTGP is straightforward. Given data (X, C), we introduce latent variables $\{Z_m\}_{m=1}^{M-1}$. Each of the corresponding trees $\{\mathcal{T}_m\}_{m=1}^{M-1}$ divides the space into an independent region set of cardinality R_m. Each tree has its own, independent, RTGP prior where the hyperparameters, parameters, and latent variable values – for fixed class index m – are generated as in (2). It is most sensible to use a constant (rather than linear) mean in each of the leaves for the RTGP latent variables in the classification context. To approximate the joint distribution of the $M-1$ TGPs, we sample with RJ-MCMC much as in Sect. 2.1. Sampling is accomplished by visiting each tree in turn. For the m^{th} class, we sequentially draw $\theta_{m,0}|\text{rest}$, $\theta_{m,r}|\text{rest}$ for each region r of R_m, $\mathcal{T}_m|\text{rest}$, and finally the latent variables $Z_{m,r}|\text{rest}$ for each r. The first three draws are the same as for the RTGP. Drawing $Z_{m,r}|\text{rest}$ is the step unique to the CTGP.

While we cannot sample directly from $Z_{m,r}|\text{rest}$ to obtain a Gibbs sampling draw, we can factorize the full conditional for some subset of $Z_{m,r}$ into the distribution of the class given the latent variables at its predictor(s) $p(C(x_I)|\{Z_{m,r}(x_I)\}_{m=1}^{M-1})$ and the distribution of the latent variable(s) given the current GP together with the other latent variables in its region $p(Z_{m,r}(x_I)|X_r, \theta, \mathcal{T}, Z_{m,r}\backslash Z_{m,r}(x_I))$. Then we can use MH and propose from the latter distribution. Here r labels the region within a particular class, and I is an index set over some of the predictors x in region r. To condense notation in (3), let $Z_I = Z_{m,r}(x_I)$ (similarly for F), and let $K_{I,I'} = K_{m,r}(x_I, x_{I'})$. Finally, $-I$ is the index set of points in region r of class m that are not in I. Then we have $Z_I|X_r, \theta, \mathcal{T}, Z_{-I} \sim \mathcal{N}_{|I|}(\hat{z}, \hat{\sigma}^2)$ with

$$\begin{aligned} \hat{z} &= F_I\tilde{\beta}_{-I} + K_{I,-I}K_{-I,-I}^{-1}(Z_{-I} - F_{-I}\tilde{\beta}_{-I}) \\ \hat{\sigma}^2 &= \sigma_r^2(\kappa_{I,I} - \kappa_{I,-I}\kappa_{-I,-I}^{-1}\kappa_{-I,I}), \qquad \kappa_{I,I'} = K_{I,I'} + \tau_r^2 F_I W F_{I'}^\top, \end{aligned} \tag{3}$$

where $\beta|X_I, Z_I, \theta, \mathcal{T} \sim \mathcal{N}_a(\tilde{\beta}_I, V_I)$ using

$$V_I^{-1} = F_I^\top K_{I,I}^{-1} F_I + W^{-1}/\tau_r^2 \qquad \tilde{\beta}_I = V(F_I^\top K_{I,I}^{-1} Z_I + W^{-1}\beta_0/\tau_r^2).$$

In this setup the prior for Z cancels with the proposal probability in the acceptance ratio. The newly proposed Z may be accepted with probability equal to the likelihood ratio:

$$A = \prod_{i \in I} \frac{\exp(-Z'_{C(x_i),r}(x_i))}{\sum_{m=1}^{M} \exp(-Z'_{m,r}(x_i))} \times \frac{\sum_{m=1}^{M} \exp(-Z_{m,r}(x_i))}{\exp(-Z_{C(x_i),r}(x_i))}.$$

We may employ a blocking scheme to increase mixing in the marginal latent Z process; however there will natural be a trade-off in block size. Proposing all components of Z_m at once leads to a small acceptance ratio and poor mixing. But proposing each component of Z_m individually may result in only small, incremental changes. An advantage of the treed partition is that it yields a natural blocking scheme for updating the latent variables. While we may block further within a leaf, this existing treed partition is a step forward from the CGP.

3 Illustrations and Empirical Results

We illustrate CTGP and compare it to CGP on real and synthetic data by making timings and calculating misclassification rates. Since the most likely class label at a particular predictor value corresponds to the largest latent variable at that predictor [via (1)], we may predict the class labels by first keeping a record of the predicted class labels at each round of the Monte Carlo run and then taking a majority vote upon completion.

3.1 2d Exponential Data

Consider the synthetic 2d exponential regression data, where the input space is $[-2, 6] \times [-2, 6]$, and the true response is given by the 2d exponential function $z(x) = x_1 \exp(-x_1^2 - x_2^2)$. To convert the real-valued outputs to classification labels we calculate the Hessian H. Then, for a particular input (x_1, x_2) we assign a class label based on the sign of the sum of the eigenvalues of $H(x_1, x_2)$, indicating the direction of concavity at that point. A function like the 2d exponential whose concavity changes more quickly in one region of the input space than in another (and is therefore well fit by an RTGP model) will similarly have class labels that change more quickly in one region than in another. The *left-hand* side of Fig. 1 shows the resulting class labels. Overlaid on the plot is the maximum *a posteriori* tree encountered in the trans-dimensional Markov chain sampling from the CTGP posterior. We trained the classifier(s) on (X, C) data obtained by a maximum entropy design of size $N = 400$ subsampled from a dense grid of 10,000 points and calculated the misclassification rate on the remaining 9,600 locations. The rate was 3.3% for CGP and 1.7% for CTGP, showing a relative improvement of roughly 50%. CTGP wins here because the relationship between response (class labels) and predictors is clearly nonstationary. The speed improvements obtained by partitioning were

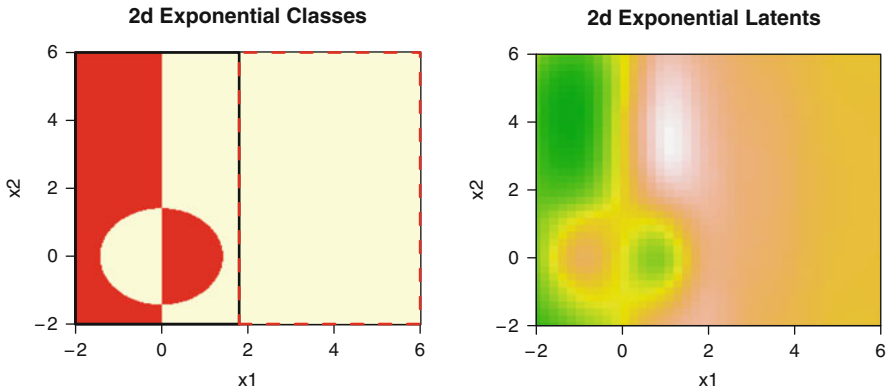


Fig. 1 2d exponential data (*left*) and mean latent variables (*right*)

even more dramatic. CGP took 21.5 h to execute 15,000 RJ-MCMC rounds, whereas CTGP took 2.0 h, an over 10-fold improvement. The *right-hand* plot in Fig. 1 shows the posterior mean of the latent variables under the CTGP model.

3.2 Classification TGP on Real Data

Consider the Credit Approval data set that may be obtained from the UCI Machine Learning database (Asuncion and Newman 2007). The set consists of 690 instances grouped into two classes: credit card application approval (+) and application failure (−). The names and values of the fifteen predictors for each instance are confidential. However, aspects of these attributes relevant to our classification task are available. E.g., we know that six inputs are continuous, and nine are categorical. Among the categorical predictors, the number of distinct categories ranges from 2 to 14. After binarization, we have a data set of six continuous and 41 binary predictors. The CGP treats these all as continuous attributes. We restrict the CTGP to form GPs only over the six continuous attributes and to apply the treed partition process on (and only on) the 41 binary attributes.

Our comparison consists of 10 separate 10-fold cross-validations for a total of 100 folds. The average misclassification rate of the CGP across these folds was 14.6% (4.0%). The CTGP offers a slight improvement with a rate of 14.2% (3.6%). More impressive is the speed-up offered by CTGP. The average CPU time per fold used by the CGP method was 5.52 h; with an average CPU time per fold of 1.62 h, the CTGP showed a more than threefold improvement.

Finally, the interpretative aspect of the CTGP is worth highlighting. For a particular run of the algorithm on the Credit Approval data, the MAP trees of different heights are shown in Fig. 2. These trees, and those for other runs, feature principal splits on the 38th binary predictor, which corresponds to the 9th two-valued categorical predictor. Therefore, the CTGP indicates, without additional work, the

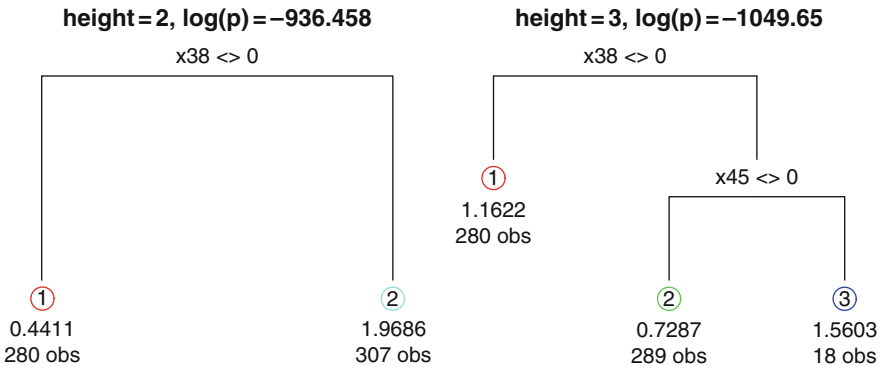


Fig. 2 Trees from CTGP on the credit approval data

significance of this variable in predicting the success of a credit card application. To extract similar information from the CGP, one would have to devise and run some additional tests – no small feat given the running time of single CGP execution.

4 Conclusion

In this paper we have illustrated how many of the benefits of the regression TGP model extend to classification. The components of TGP, i.e. treed models and GPs, have separately long enjoyed success in application to classification problems. In the case of the GP, $M - 1$ processes are used as a prior for latent variables which encode the classes via a softmax function. While this is a powerful method which typically offers improvements over simpler approaches (including treed models), drawbacks include an implicit assumption of stationarity and slow evaluation due to repeated large matrix decompositions. In contrast, the treed methods provide a thrifty divide-and-conquer approach. The combined tree and GP approach provides a classification model that is speedy, interpretable, and highly accurate, combining the strengths of GP and treed models for classification.

References

Asuncion, A., & Newman, D. (2007). UCI Machine Learning Repository.

Breiman, L., Friedman, J. H., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Chipman, H., George, E., & McCulloch, R. (1998). Bayesian CART model search (with discussion). *Journal of the American State Association, 93*, 935–960.

Chipman, H., George, E., & McCulloch, R. (2002). Bayesian treed models. *Machine Learning, 48*, 303–324.

Gramacy, R. B., & Lee, H. K. H. (2008). Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association, 103*, 1119–1130.

Neal, R. M. (1998). Regression and classification using Gaussian process priors (with discussion). In J. M. Bernardo (Ed.), *Bayesian Statistics 6* (pp. 476–501). Oxford: Oxford University Press.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Cambridge, MA: The MIT Press.

Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B, Methodological, 59*, 731–758.

Stein, M. L. (1999). *Interpolation of Spatial Data*. New York, NY: Springer.

Ridgeline Plot and Clusterwise Stability as Tools for Merging Gaussian Mixture Components

Christian Hennig

Abstract The problem of merging Gaussian mixture components is discussed in situations where a Gaussian mixture is fitted but the mixture components are not separated enough from each other to interpret them as "clusters". Two methods are introduced, corresponding to two different "cluster concepts" (separation by gaps and "data patterns"). A visualisation of the modality of a density of a mixture of two Gaussians is proposed and the stability of the unmerged Gaussian mixture is compared to that of clusterings obtained by merging components.

1 Introduction

The Gaussian mixture model is often used for cluster analysis (Fraley and Raftery 2002). $I\!R^p$-valued observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are modelled as i.i.d. with density

$$f(\mathbf{x}) = \sum_{j=1}^{s} \pi_j \varphi_{\mathbf{a}_j, \Sigma_j}(\mathbf{x}), \tag{1}$$

where $\pi_j > 0 \; \forall j$, $\sum_{j=1}^{s} \pi_j = 1$, $\varphi_{\mathbf{a},\Sigma}$ is the density of the p-dimensional Gaussian distribution $\mathcal{N}(\mathbf{a}, \Sigma)$ with mean vector $\mathbf{a}$ and covariance matrix Σ. Given a fixed s, the parameters can be estimated by Maximum Likelihood using the EM algorithm. The data points can then be classified to the mixture components by maximizing the estimated a posteriori probability that $\mathbf{x}_i$ was generated by mixture component j,

$$\hat{P}(\gamma_i = j | \mathbf{x}_i = \mathbf{x}) = \frac{\hat{\pi}_j \varphi_{\hat{\mathbf{a}}_j, \hat{\Sigma}_j}(\mathbf{x}))}{\sum_{l=1}^{s} \hat{\pi}_l \varphi_{\hat{\mathbf{a}}_l, \hat{\Sigma}_l}(\mathbf{x})}, \tag{2}$$

where γ_i is defined by the two-step version of the mixture model where

C. Hennig
Department of Statistical Science, UCL, Gower Street, London WC1E 6BT, United Kingdom
e-mail: chrish@stats.ucl.ac.uk

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_11,

$$P(\gamma_i = j) = \pi_j,\ \mathbf{x}_i|(\gamma_i = j) \sim \varphi_{\mathbf{a}_j,\Sigma_j},\ i = 1,\ldots,n,\ \text{i.i.d.} \tag{3}$$

Estimators are denoted by "hats". A standard method to estimate the number of components s is the Bayesian Information Criterion (BIC). This can also be used to estimate suitable constraints on the covariance matrices (Fraley and Raftery 2002). For the present paper, the Gaussian mixture model has been fitted using the default options of the add-on package MCLUST version 3 (Fraley and Raftery 2006) of the statistical software R (www.R-project.org).

In cluster analysis usually every mixture component is interpreted as corresponding to a cluster (which generally is a subset of the data with the interpretation that its members belong together in some sense), and pointwise maximization of (2) defines the clustering. However, this is often not justified. Some mixtures of more than one Gaussian distribution are unimodal, and in reality, model assumptions are never precisely fulfilled and Gaussian mixtures are a very flexible tool to fit all kinds of densities. But this means that a population that can be interpreted as "homogeneous" could be fitted by a mixture of more than one Gaussian mixture component.

Figure 1 shows an artificial dataset generated from a mixture of two uniform distributions. MCLUST with its default settings for estimating Gaussian mixtures estimates $s = 3$ for this dataset. Note that the "correct" number of clusters is not well defined and one may see two "true" clusters here if "distinguishable patterns" are interpreted to be clusters or one "true" cluster if clusters are associated with density modes or are required to be separated by gaps. Depending on the cluster concept of interest in a given application, clusters may be associated with modes,

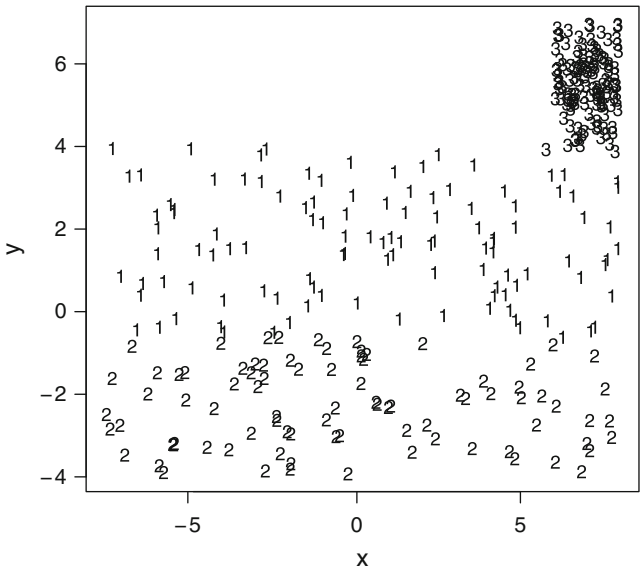


Fig. 1 Data from unimodal mixture of uniforms with 3-cluster solution by MCLUST

or with some other kinds of clear patterns in the data. However, three is hardly justifiable as the number of clusters here.

In the present paper, two methods are introduced that can be used to decide whether and which Gaussian mixture components should be merged. A method based on detecting density gaps is introduced in Sect. 2. A method based on estimating misclassification probabilities is introduced in Sect. 3. Detailed background for both methods along with some alternatives is given in Hennig (2010), so they are only explained very briefly here. The following two tools are introduced exclusively in the present paper: the idea of Sect. 2 can be used to define "ridgeline plots" that show how strongly mixture components are separated, and in Sect. 4 the clusterwise stability assessment method introduced in Hennig (2007) is proposed to compare the stability of the MCLUST and the merged clustering solution.

A real dataset from musicology is analysed in Sect. 5, and Sect. 6 concludes the paper.

Further methods for merging Gaussian mixture components are given in Tantrum et al. (2003) and Li (2004). Both of them are discussed in Hennig (2010).

2 The Ridgeline Method

In Ray and Lindsay (2005) it is shown that for any mixture f of s Gaussian distributions on $\mathbb{R}^p$ there is an $(s-1)$-dimensional manifold of $\mathbb{R}^p$ so that all local maxima and minima of f lie on this manifold.

For $s = 2$, this manifold is defined by the so-called "ridgeline",

$$\mathbf{x}^*(\alpha) = [(1-\alpha)\Sigma_1^{-1} + \alpha\Sigma_2^{-1}]^{-1}[(1-\alpha)\Sigma_1^{-1}\mathbf{a}_1 + \alpha\Sigma_2^{-1}\mathbf{a}_2], \tag{4}$$

and all density extrema (and therefore all modes, which may be more than two in some situations) can be found by looking up the density at $\mathbf{x}^*(\alpha)$, $\alpha \in [0,1]$. The following algorithm ("ridgeline method") can be used to merge mixture components in order to merge unimodal groups of Gaussian components:

1. Choose a tuning constant $r^* < 1$.
2. Start with all components of the initially estimated Gaussian mixture (for example obtained by MCLUST's EM-implementation) as current clusters.
3. Using the mean vectors and covariance matrices of the current clusters, for any pair of current clusters compute, from (4), $r = \frac{\min_{0\leq\alpha\leq 1} f(\mathbf{x}^*(\alpha))}{m_2}$, where m_2 denotes the second largest mode of the mixture density restricted to the current pair of components; let $r = 1$ if there is only one mode.
4. If $r < r^*$ for all pairs of current clusters, use the current clustering as the final one.
5. Otherwise, merge the pair of current clusters with maximum r and go to step 3.

Following Hennig (2010), $r^* = 0.2$ is used here. Note that it is not advisable to demand strict unimodality ($r^* = 1$), because the probability is high that MCLUST estimates multimodal mixtures even in unimodal situations. For example, for the

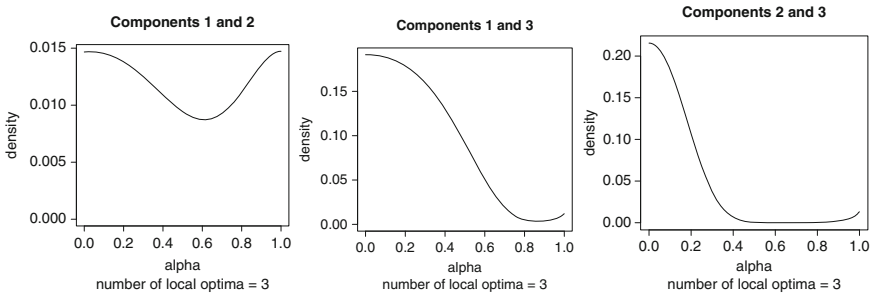


Fig. 2 Ridgeline plot $f(\mathbf{x}^*(\alpha))$ for the three Gaussian components estimated by MCLUST for the data in Fig. 1

(unimodal) dataset in Fig. 1, the first two components are merged at $r = 0.58$ and their union is merged with the third one at $r = 0.21$.

The separation of the estimated Gaussian mixture components can be visualised by plotting the ridgeline density $f(\mathbf{x}^*(\alpha))$ vs. α. The results for the dataset in Fig. 1 are shown in Fig. 2.

3 A Method Based on Misclassification Probabilities

The basic idea if the method based on directly estimated misclassification probabilities (DEMP) is that, alternatively to a modality based cluster concept, mixture components could be merged if the misclassification probability between them is high.

Let γ_k^* denote the membership indicator of observation k in a Gaussian mixture model, but where a cluster may be identified with a single mixture component [in which case $\gamma_k^* = \gamma_k$ in (3)], or a mixture of several Gaussian components. π_i^* denotes the prior probability for component i. $\tilde{\gamma}_k^*$ is the mixture component to which the point is classified by the Bayes rule with true parameters. $1(\bullet)$ denotes the indicator function.

Misclassification probabilities $p_{ij} = P(\tilde{\gamma}_k^* = i|\gamma_k^* = j) = \frac{P(\tilde{\gamma}_k^*=i,\ \gamma_k^*=j)}{\pi_j^*}$ between components of a mixture distribution can be estimated directly from the results of the EM algorithm for Gaussian mixtures. Estimators $\hat{\pi}_i^*$ can be obtained straightforward by summing up the $\hat{\pi}_m$ of the Gaussian member components of the mixture of mixtures i. Note that

$$\hat{P}(\tilde{\gamma}_1^* = i,\ \gamma_1^* = j) = \frac{1}{n}\sum_{h=1}^{n} \hat{P}(\gamma_h^* = j|\mathbf{x}_h)1(\hat{\gamma}_h^* = i) \tag{5}$$

is a consistent estimator of $P(\tilde{\gamma}_1^* = i,\ \gamma_1^* = j)$, where $\hat{\gamma}_h^*$ denotes the estimated maximum a posteriori classification of data point $\mathbf{x}_h$ (i.e., maximising $\hat{P}(\gamma_h^* = j|\mathbf{x}_h)$), which estimates $\tilde{\gamma}_h^*$.

Therefore,

$$\hat{p}_{ij} = \frac{\hat{P}(\tilde{\gamma}_1^* = i,\ \gamma_1^* = j)}{\hat{\pi}_j^*}$$

is a consistent estimator of p_{ij}. This works regardless of whether the mixture components are Gaussian distributions or mixtures of Gaussians. Here is the DEMP method.

1. Choose a tuning constant $q^* < 1$.
2. Start with all components of the initially estimated Gaussian mixture as current clusters.
3. Compute $q = \max(\hat{p}_{ij}, \hat{p}_{ji})$ for all pairs of current clusters.
4. If $q < q^*$ for all pairs of current clusters, use the current clustering as the final one.
5. Otherwise, merge the pair of current clusters with maximum q and go to step 3.

$q^* = 0.025$ is used here (Hennig 2010).

For the data in Fig. 1, DEMP merges components 1 and 2 at $q = 0.078$. For the cluster of these two and component 3, $q = 0.01$, so with $q^* = 0.025$, two clusters are found, namely mixture components 1/2 together and 3. This makes sense if clusters refer to "patterns" in the data but are not required to be separated by gaps.

4 Bootstrap Stability Assessment

In Hennig (2007), the following idea has been introduced for checking the stability of a cluster in a given clustering $\{C_1, \ldots, C_s\}$ (Hennig 2007; applies to quite general clustering methods).

- Draw B nonparametric bootstrap samples from the dataset [it is advisable to discard the copies of points drawn more than once; further schemes to generate datasets are discussed in Hennig (2007)].
- Cluster the bootstrapped datasets by the same method that was used for the original dataset.
- For every cluster in the original dataset, find the most similar one in every bootstrapped dataset. Similarity is measured according to the Jaccard similarity between sets C, D : $j(C, D) = \frac{|C \cap D|}{|C \cup D|}$.
- For every cluster C_i, $\bar{j}_i = \frac{1}{B}\sum_{k=1}^{B} j_{max,k}(C_i)$, where $j_{max,k}(C_i)$ is the Jaccard similarity of C_i to the most similar cluster in bootstrap sample k.

The Jaccard similarity is between 0 and 1. It makes sense to consider clusters with $\bar{j}_i < 0.5$ as "dissolved" (Hennig 2007) and a meaningful stable cluster should have $\bar{j}_i \gg 0.5$, better above 0.7 or 0.8 (though not every stable cluster is meaningful).

In the given situation it is interesting to apply the idea to the MCLUST clusterings and compare them to the stability achieved by the clusters yielded by the merging

methods, i.e., to consider "do MCLUST first and apply ridgeline or DEMP method to the solution" as a clustering method in its own right.

For the dataset in Fig. 1 and the MCLUST solution, $\bar{j}_1 = 0.48$, $\bar{j}_2 = 0.52$, $\bar{j}_3 = 0.96$, so the first two clusters are obviously unstable. The ridgeline method yields $\bar{j}_{123} = 0.87$ for the only remaining cluster, which means that in some situations it ends up with more than one cluster (for merging methods, the lower index of $\bar{j}$ refers to the original Gaussian components belonging to the merged cluster). DEMP yields $\bar{j}_{12} = 0.94$, $\bar{j}_3 = 0.98$, which confirms that this is a very stable solution.

5 Real Data Example: Clustering Melody Contours

The dataset analysed here consists of approximations of the contours of 989 melody phrases taken from commercial pop songs by polynomials of degree 5, as discussed in Frieler et al. (in press). The dataset was provided by D. Müllensiefen. Due to a lower intrinsic dimensionality of the dataset, only four coefficients (a4, a3, a1 and a0, indicating the coefficients belonging to the terms x^4, x^3, x^1, x^0 of the polynomials) were used as variables, and most of the information distinguishing the seven clusters obtained by MCLUST can be seen in the scatterplot of a4 and a0, see Fig. 3. Apart from a strong concentration around the value 0 of the first variable (which is made up by points of two components, no. 1 and 4, in the MCLUST solution, see right side of Fig. 3), no clear patterns can be seen.

The stability values for the MCLUST solution are: $\bar{j}_1 = 0.75$, $\bar{j}_2 = 0.33$, $\bar{j}_3 = 0.42$, $\bar{j}_4 = 0.42$, $\bar{j}_5 = 0.27$, $\bar{j}_6 = 0.24$, $\bar{j}_7 = 0.44$, so only the first component is reasonably stable. This in itself is very useful for interpreting the clustering, even before having done any component merging.

Some ridgeline plots are given in Fig. 4. Note that the gap between components 2 and 5 on the lower left side is by far the deepest for any pair of components in this dataset, which indicates that the 2-d plots in Fig. 3 do not miss any strong separation

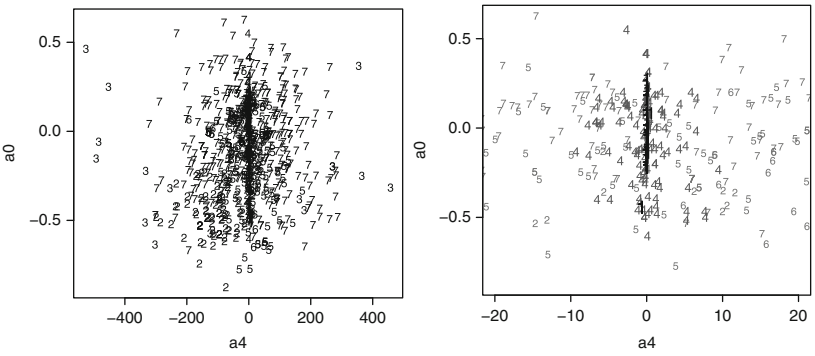


Fig. 3 Variables a4 and a0 of melody data with clusters found by MCLUST (*right side*: magnified version of the central area; what looks like a "line" along a4 ≈ 0 is component 1)

between any two clusters in 4-d space (particularly taking into account that even in mixtures without clear gaps, gaps are expected to occur in ridgeline ratio plots between "non-neighbouring" components, see Fig. 2). Some pairs of components do not yield unimodal mixtures, but one of the modes is usually very weak, as for example for the two component pairs on the right side of Fig. 4 (it is hardly visible that the mixture of components 2 and 4 is bimodal but the density goes up a tiny little bit approaching $\alpha = 1$). About half of the pairs of components yield unimodal mixtures such as on the upper left side of Fig. 4.

The ridgeline method merges all clusters, justified by the fact that there are no clear gaps. This is very stable ($\bar{j}_{1234567} = 1$). However, demanding estimated unimodality by using $r^* = 1$ in Sect. 2 merges all components except of component 2 with stability $\bar{j}_2 = 0.12$, indicating again that $r^* = 1$ is not a good idea. DEMP yields three clusters by leaving components 1 and 4 unmerged and merging all the others. These clusters are obviously not separated by gaps, but correspond to visible patterns in Fig. 3. The stabilities are $\bar{j}_1 = 0.74$, $\bar{j}_4 = 0.47$, $\bar{j}_{23567} = 0.91$. This

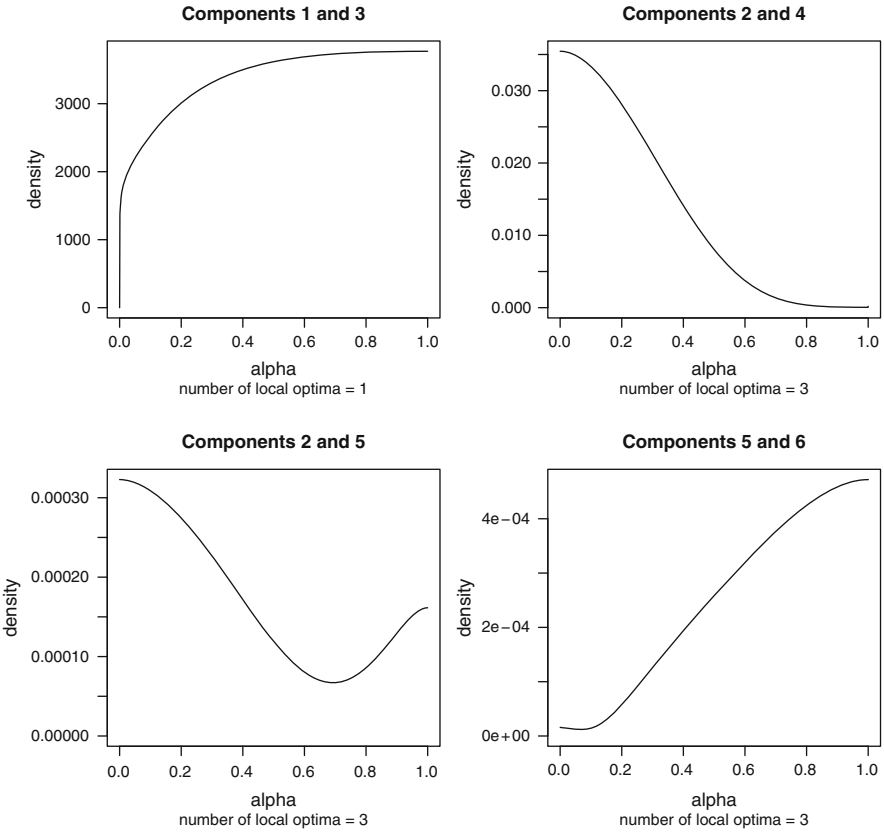


Fig. 4 Selected ridgeline plots for melody data

indicates that it makes sense to distinguish component 1 from the union of components 2, 3, 5, 6, 7 as a stable pattern. Component 4, which lies "between" the other two clusters, cannot be so clearly distinguished from them. This corresponds nicely with the visual impression from Fig. 3. In terms of the melodic phrases, there are no groups of phrases that can really be separated from the others by "gaps", but there is a core pattern of phrases (component 1 and to some extent 4) that can be interpreted as having in common a value of about zero for the fourth degree coefficient (a4) of the contour approximating polynomial, which means that no steep increase/decrease (or decrease/increase) combinations occur in the melody contour.

6 Conclusion

The problem of merging Gaussian mixture components to find more meaningful clusters cannot be uniquely solved. Solutions always depend on what kind of clusters the researcher is looking for. For example, clusters can be defined by gaps (rather corresponding to the ridgeline method) or by "patterns" (rather corresponding to the DEMP method). Visualisation of the separation of mixture components and assessment of the stability of clusters can help with the decision whether some of the original mixture components should be merged, and with how the results are to be interpreted. Further methods for merging and visualisation, details, examples and comparisons (including some situations in which DEMP merges stronger than the ridgeline method as opposed to the two examples here) are given in Hennig (2010).

References

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97*, 611–631.

Fraley, C., & Raftery, A. E. (2006). MCLUST Version 3 for R: normal mixture modeling and model-based clustering. Technical Report no. 504, Department of Statistics, University of Washington.

Frieler, K., Müllensiefen, D., & Riedemann, F. (in press). Statistical search for melodic prototypes. In T. Klouche (Ed.). *Conference on Mathematics in Music*. Berlin: Staatliches Institut für Musikforschung.

Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics and Data Analysis, 52*, 258–271.

Hennig, C. (2010). Methods for merging Gaussian mixture components. *Advances in Data Analysis and Classification, 4*, 3–34.

Li, J. (2004). Clustering based on a multilayer mixture model. *Journal of Computational and Graphical Statistics, 14*, 547–568.

Ray, S., & Lindsay, B. G. (2005). The topography of multivariate normal mixtures. *Annals of Statistics, 33*, 2042–2065.

Tantrum, J., Murua, A., & Stuetzle, W. (2003). Assessment and pruning of hierarchical model based clustering. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, D.C.*, 197–205.

Clustering with Confidence: A Low-Dimensional Binning Approach

Rebecca Nugent and Werner Stuetzle

Abstract We present a plug-in method for estimating the cluster tree of a density. The method takes advantage of the ability to exactly compute the level sets of a piecewise constant density estimate. We then introduce *clustering with confidence*, an automatic pruning procedure that assesses significance of splits (and so clusters) in the cluster tree; the only user input required is the desired confidence level.

1 Introduction

The goal of clustering is to identify distinct groups in a data set and assign a group label to each observation. Ideally, we would be able to find the number of groups as well as where each group lies in the feature space with minimal input from the user. To cast clustering as a statistical problem, we regard the data, $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\} \in \mathbf{R^p}$, as a sample from some unknown population density $p(\mathbf{x})$. There are two statistical approaches. While parametric (model-based) clustering assumes the data have been generated by a finite mixture of g underlying parametric probability distributions $p_g(\mathbf{x})$ (often multivariate Gaussian) (Fraley and Raftery 1998; McLachlan and Basford 1988), the nonparametric approach assumes a correspondence between groups in the data and modes of the density $p(\mathbf{x})$. Wishart first advocated searching for modes as manifestations of the presence of groups (Wishart 1969); nonparametric clustering should be able to "resolve distinct data modes, independently of their shape and variance". Hartigan expanded this idea and made it more precise (Hartigan 1975, 1981).

Define a level set $L(\lambda; p)$ of a density p at level λ as the subset of the feature space for which the density exceeds λ: $L(\lambda; p) = \{\mathbf{x} | p(\mathbf{x}) > \lambda\}$. Its connected components are the maximally connected subsets of a level set. For any two connected components A and B, possibly at different levels, either $A \subset B$, $B \subset A$, or

R. Nugent (✉)
Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: rnugent@stat.cmu.edu

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_12,

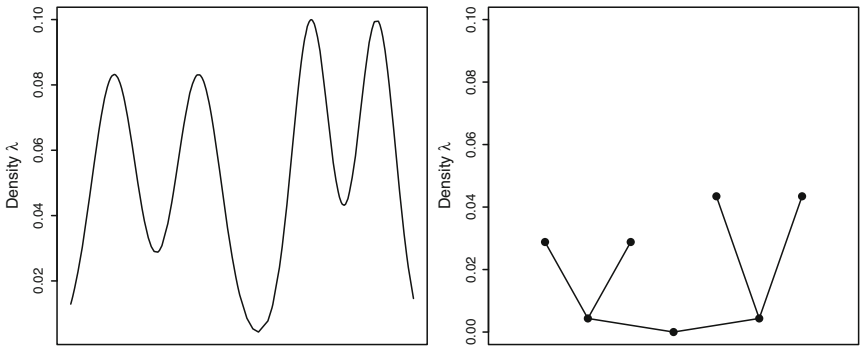


Fig. 1 (a) Density with four modes; (b) cluster tree with three splits/four leaves

$A \cap B = \emptyset$. This hierarchical structure of the level sets is summarized by the *cluster tree* of p.

The cluster tree is easiest to define recursively (Stuetzle 2003). Each node N of the tree represents a subset $D(N)$ of the support $L(0; p)$ of p and is associated with a density level $\lambda(N)$. The root node represents the entire support of p and is associated with density level $\lambda(N) = 0$. To determine the daughters of a node, we find the lowest level λ_d for which $L(\lambda_d; p) \bigcap D(N)$ has two or more connected components. If no such λ_d exists, then $D(N)$ is a mode of the density, and N is a leaf of the tree. Otherwise, let $C_1, C_2, \ldots, C_n$ be the connected components of $L(\lambda_d; p) \bigcap D(N)$. If $n = 2$, we create two daughter nodes at level λ_d, one for each connected component; we then apply the procedure recursively to each daughter node. If $n > 2$, we create two connected components C_1 and $C_2 \bigcup C_3 ... \bigcup C_n$ and their respective daughter nodes and then recurse. We call the regions $D(N)$ the "high density clusters" of p. This recursive binary tree also can accommodate level sets with more than two connected components with repeated splits at the same height. Figure 1 shows a univariate density with four modes and the corresponding cluster tree with initial split at $\lambda = 0.0044$ and subsequent splits at $\lambda = 0.0288, 0.0434$. Estimating the cluster tree is a fundamental goal of nonparametric cluster analysis.

There are several previously suggested clustering methods based on level sets and other level set estimation procedures. In general, they are heuristic in nature or require subjective decisions from the user (Wishart 1969; Walther 1997; Cuevas et al. 2000, 2001; Stuetzle 2003; Klemelä 2004).

2 Cluster Trees: Piecewise Constant Density Estimates

We can estimate the cluster tree of a density p by the cluster tree of a density estimate $\hat{p}$. However, for most density estimates, computing the cluster tree is a difficult problem; there is no obvious method for computing and representing the level sets.

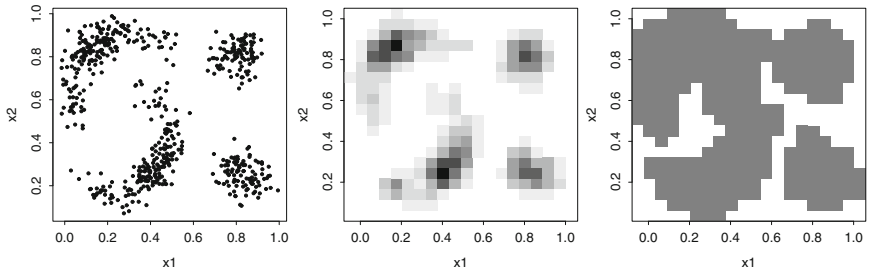


Fig. 2 (a) Four well-separated groups; (b) BKDE, 20×20 grid (c) $L(0.00016;\text{BKDE})$

Exceptions are density estimates that are piecewise constant over (hyper-)rectangles. Let $B_1, B_2, \ldots, B_N$ be the rectangles, and let $\hat{p}_i$ be the estimated density for B_i. Then $L(\lambda; \hat{p}) = \bigcup_{\hat{p}_i > \lambda} B_i$. If the dimension is low enough, any density estimate can be reasonably binned. Here we use the Binned Kernel Density Estimate (BKDE $= \hat{p}$), a binned approximation to an ordinary kernel density estimate, on a grid we can computationally afford (Wand and Jones 1995; Hall and Wand 1996). We use 10-fold cross validation to estimate the bandwidth h.

Figure 2a has four well-separated groups, two curvilinear and two spherical; a grey-scale heat map of a BKDE on a 20×20 grid ($h = 0.0244$) is in Fig. 2b. Figure 2c shows the level set $L(0.00016; \hat{p})$; since we have a split, we would create two daughter nodes at this height. When a cluster tree node N has been split into daughters N_l, N_r, the high density clusters $D(N_l)$, $D(N_r)$, also referred to as the cluster "cores", do not necessarily form a partition of $D(N)$. We refer to the bins (and their observations) in $D(N)\backslash(D(N_l) \cup D(N_r))$, e.g. the white bins in Fig. 2c, as the "fluff". We assign each fluff bin B to N_r if the Manhattan distance $d_M(B, D(N_r)) = \|B - D(N_r)\|_1$ is less than $d_M(B, D(N_l))$. If $d_M(B, D(N_r)) > d_M(B, D(N_l))$, then B is assigned to N_l. In case of ties, the algorithm arbitrarily chooses an assignment. The cluster cores and fluff represented by the leaves of the cluster tree form a partition of the support of $\hat{p}$ and a corresponding partition of the observations. The same is true for every subtree of the cluster tree.

During cluster tree construction, $L(\lambda_d; \hat{p}) \bigcap D(N)$ changes structure only when the level λ_d is equal to the next higher value of $\hat{p}(B_i)$ for one or more bins B_i in $D(N)$. We compute the cluster tree of $\hat{p}$ by "stepping through" the bins' sorted unique density estimate values; every increase in level λ_d then corresponds to the removal of one or more bins from the level set $L(\lambda_d; \hat{p})$. We represent $L(\lambda_d; \hat{p}) \cap D(N)$ as an adjacency graph G where the vertices $B_i \in L(\lambda_d; \hat{p}) \cap D(N)$ are connected by an edge if they share a lower-dimensional face. Finding its connected components is a standard graph problem (Robert 2002).

Figure 3a shows the BKDE's cluster tree; the corresponding cluster assignments and partitioned feature space are in Fig. 3b,c. The cluster tree indicates the BKDE has nine modes. The first split at $\lambda = 1.9 \cdot 10^{-16}$ and the mode around (1, 0.5)

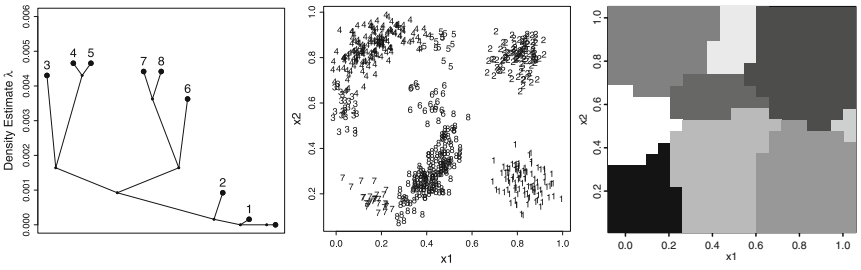


Fig. 3 (a) BKDE cluster tree; (b) cluster assignments; (c) partitioned feature space

in Fig. 3c are artifacts of $\hat{p}$; no observations are assigned to the resulting daughter node. The remaining eight leaves correspond to (a subset of) one of the groups.

We briefly compare these results to common clustering methods (not described here) by comparing the estimated clusters to the true groups using the Adjusted Rand Index (Hubert and Arabie 1985). The ARI is a common measure of agreement between two partitions. Its expected value is zero under random partitioning with a maximum value of one; larger values indicate better agreement. Using the total within-cluster sum of squares criterion, k-means (Mardia et al. 1979) selects five or six clusters with ARI = 0.803, 0.673; for $k = 4$, ARI = 0.924. Model-based clustering (MBC, Fraley and Raftery 1998; McLachlan and Basford 1988) with an unconstrained covariance structure chose ten clusters (ARI = 0.534). The BKDE cluster tree on a 20×20 grid performed comparably ($k = 8$; ARI = 0.781); a 15×15 grid performed slightly better ($k = 6$; ARI = 0.865). The groups are well-separated; however, the two curvilinear groups give an increased number of clusters (k-means, MBC, and the cluster tree). Single, complete, and average hierarchical linkage methods (Mardia et al. 1979) gave perfect agreement given knowing the true number of groups in advance.

For both grid choices, the cluster tree overestimated the number of groups (8,6). Figure 3 illustrates this problem in the approach. While the cluster tree is accurate for the given density estimate, the inherent noise in the density estimate results in spurious modes not corresponding to groups in the underlying population. In our example, the procedure identified the four original groups (post the modal artifact) but erroneously continued splitting the clusters. The corresponding branches of the cluster tree need to be pruned.

3 Clustering with Confidence

We propose a bootstrap-based automatic pruning procedure that finds simultaneous upper and lower $(1 - \alpha)$ confidence sets for each level set. During cluster tree construction, only splits indicated as significant by the bootstrap confidence sets are taken to signal multi-modality. Spurious modes are discarded during estimation; the only user decision is the confidence level.

3.1 Bootstrap Confidence Sets for Level Sets

We define upper confidence sets (UCS) to be of the form $L^u(\lambda;\hat{p}) = L(\lambda - \delta^u_\lambda;\hat{p})$ and lower confidence sets (LCS) of form $L^l(\lambda;\hat{p}) = L(\lambda + \delta^l_\lambda;\hat{p})$ with $\delta^u_\lambda, \delta^l_\lambda > 0$. By construction, LCS $= L^l(\lambda;\hat{p}) \subseteq L(\lambda;\hat{p}) \subseteq L^u(\lambda;\hat{p}) =$ UCS.

Let $\hat{p}^*_1, \hat{p}^*_2, \ldots, \hat{p}^*_m$ be the density estimates for m bootstrap samples of size n drawn with replacement from the original sample. We call a pair $(L(\lambda - \delta^u_\lambda;\hat{p}), L(\lambda + \delta^l_\lambda;\hat{p}))$ a non-simultaneous $(1-\alpha)$ confidence set for $L(\lambda;p)$ if for $100\cdot(1-\alpha)\%$ of the bootstrap density estimates $\hat{p}^*_i$, the upper confidence set $L(\lambda - \delta^u_\lambda;\hat{p})$ *contains* $L(\lambda;\hat{p}^*_i)$, and the lower confidence set $L(\lambda+\delta^l_\lambda;\hat{p})$ *is contained in* $L(\lambda;\hat{p}^*_i)$:

$$P_{boot}\{L(\lambda + \delta^l_\lambda;\hat{p}) \subseteq L(\lambda;\hat{p}^*_i) \subseteq L(\lambda - \delta^u_\lambda;\hat{p})\} \geq 1-\alpha.$$

Here is one method to determine $\delta^u_\lambda, \delta^l_\lambda$ (Buja 2002). For each bootstrap sample $\hat{p}^*_i$ and each of the finitely many levels of $\hat{p}$, find the smallest $\delta^u_\lambda(i)$ such that $L(\lambda;\hat{p}^*_i) \subseteq L(\lambda - \delta^u_\lambda(i);\hat{p})$ and the smallest $\delta^l_\lambda(i)$ such that $L(\lambda + \delta^l_\lambda(i);\hat{p}) \subseteq L(\lambda;\hat{p}^*_i)$. Choose $\delta^u_\lambda = (1-\frac{\alpha}{2})$ quantile of the $\delta^u_\lambda(i)$ and $\delta^l_\lambda = (1-\frac{\alpha}{2})$ quantile of the $\delta^l_\lambda(i)$. By construction, the pair $(L(\lambda - \delta^u_\lambda;\hat{p}), L(\lambda + \delta^l_\lambda;\hat{p}))$ is a $(1-\alpha)$ non-simultaneous confidence set for $L(\lambda;p)$. To get confidence sets for all λ occurring as values of $\hat{p}$ with simultaneous coverage probability $1-\alpha$, we simply increase the coverage level of the individual sets until the desired level of simultaneous coverage is reached. Note that the actual upper and lower confidence sets for $L(\lambda;p)$ are the level sets $(L(\lambda - \delta^u_\lambda;\hat{p}), L(\lambda + \delta^l_\lambda;\hat{p}))$ respectively for $\hat{p}$. The bootstrap is used only to find $\delta^u_\lambda, \delta^l_\lambda$.

3.2 Constructing the Cluster Tree

After finding $\delta^u_\lambda, \delta^l_\lambda$ for all λ, we incorporate the bootstrap confidence sets into the cluster tree construction by only allowing splits at heights λ for which the corresponding bootstrap confidence set $(L^l(\lambda;\hat{p}), L^u(\lambda;\hat{p}))$ gives strong evidence of a split. We use a similar recursive procedure to that in Sect. 2. The root node represents the entire support of $\hat{p}$ and is associated with density level $\lambda(N) = 0$. To determine the daughters of a node, we find the lowest level λ_d for which a) $L^l(\lambda_d;\hat{p}) \bigcap D(N)$ has two or more connected components that b) are disconnected in $L^u(\lambda_d;\hat{p}) \bigcap D(N)$. Condition (a) indicates that the underlying density p has two peaks above height λ; condition (b) indicates that the two peaks are separated by a valley dipping below height λ. Satisfying both conditions indicates a split at height λ. If no such λ_d exists, N is a leaf of the tree. Otherwise, let C^l_1, C^l_2 be two connected components of $L^l(\lambda_d;\hat{p}) \bigcap D(N)$ that are disconnected in $L^u(\lambda_d;\hat{p}) \bigcap D(N)$. Let C^u_1 and C^u_2 be the connected components of $L^u(\lambda_d;\hat{p}) \bigcap D(N)$ from the possible $C^u_1, C^u_2, \ldots, C^u_n$ that contain C^l_1 and C^l_2

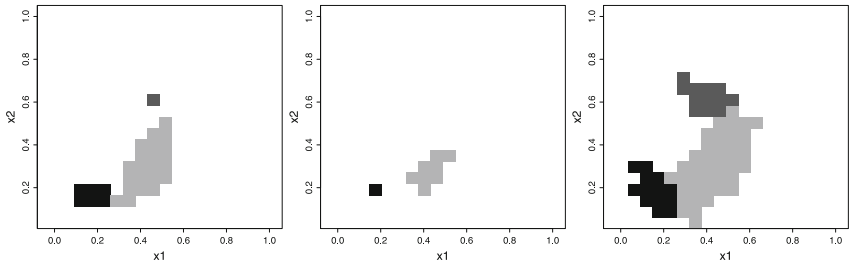


Fig. 4 (a)$L(0.0036; \hat{p})$; (b) LCS, $\delta_\lambda^l = 0.0063$; (c) UCS, $\delta_\lambda^u = 0.0028$

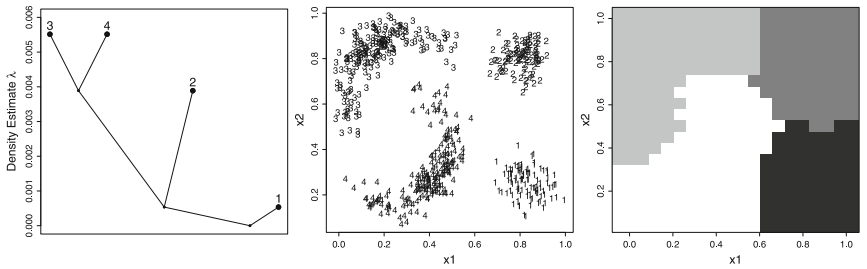


Fig. 5 (a) 95% confidence cluster tree (b) clusters; (c) partitioned feature space

respectively. If $n = 2$, we create two daughter nodes at level λ_d for C_1^u and C_2^u and, to each, apply the procedure recursively. If $n > 2$, we create two connected components C_1^u and $C_2^u \bigcup C_3^u ... \bigcup C_n^u$ and respective daughter nodes and recurse.

We return to the split at $\lambda = 0.0036$, the first split that breaks the lower left curvilinear group into two clusters (Fig. 3a: 6 and 7,8). Figure 4 shows the bootstrap confidence set ($\alpha = 0.05$) for this level set. The original level set $L(0.0036; \hat{p})$ is in Fig. 4a (grey-scale corresponds to Fig. 3 final leaf). The LCS is found to be $\delta_\lambda^l = 0.0063$ higher, i.e. $L(0.0099; \hat{p})$ (Fig. 4b). The UCS is found to be $\delta_\lambda^u = 0.0028$ lower, i.e. $L(0.0008; \hat{p})$ (Fig. 4c). At $\lambda = 0.0008$, the UCS does not have two connected components (no valley). Moreover, even though the LCS does have two connected components, they do not correspond to the two connected components in $L(0.0036; \hat{p})$. We do not have evidence of a significant split and so do not create daughter nodes at this level.

Clustering with Confidence (CWC) with $\alpha = 0.05$ generates the cluster tree and data/feature space partitions in Fig. 5. The cluster tree's significant splits have identified the four original groups as significant clusters (ARI = 1). No other smaller clusters (or modal artifacts) are found. Note that the split heights are higher than the corresponding split heights in the cluster tree in Fig. 3. The CWC procedure required stronger evidence for a split than was available at the lower levels. It performed more favorably than k-means or model-based clustering and provided a measure of confidence for the clusters.

4 Example: "Automatic Gating" in Flow Cytometry

The algorithms presented could be used for any number of dimensions but are more tractable for lower dimensions. For easier visualization of the results, we present a real two-dimensional application from molecular biology. We comment on higher dimensionality in the summary and future work section.

Flow cytometry is a technique for examining and sorting tagged mRNA molecules in a cell population. Each cell's fluorescence level (corresponding to, e.g., gene expression level) is recorded as particles pass in front of a single wavelength laser. We are interested in discovering groups of cells with high fluorescence levels for multiple channels or groups of cells that have different levels across channels. A common identification method is "gating" or subgroup extraction from two-dimensional plots of measurements on two channels. Most commonly, these subgroups are identified by eyeballing the graphs. Clustering techniques would allow for more statistically motivated subgroup identification (Lo et al. 2008).

We have 1,545 flow cytometry measurements on two fluorescence markers (anti-BrdU, binding dye 7-AAD) applied to Rituximab, a therapeutic monoclonal antibody, in a drug-screening project designed to identify agents to enhance its anti-lymphoma activity (Lo et al. 2009). Figure 6a shows the cluster tree (BKDE 15×15; $h = 21.834$); the cluster assignments as well as whether or not the observations are part of a cluster "core" (larger labels) are in Fig. 6b. The cluster tree has 12 leaves (8 clusters, 4 modal artifacts). The core sizes give some evidence as to their eventual significance. For example, cluster 1's core near (500, 1,000) contains one observation; we would not expect cluster 1 to remain in the confidence cluster tree for any reasonable α.

We use CWC to construct a confidence cluster tree for $\alpha = 0.10$; we are at least 90% confident in the generated clusters (Fig. 6c). All modal artifacts have been removed; the smaller clusters are merged into two larger clusters with cores at (200, 200), (700, 300). Note that the right cluster is a combination of the mid to high 7-AAD clusters in Fig. 6b. CWC did not find enough evidence to warrant splitting this larger cluster further into subgroups.

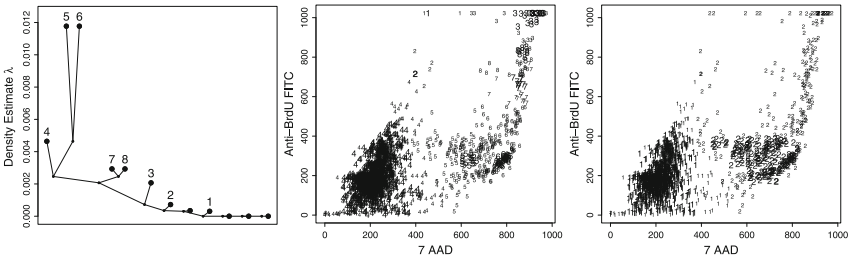


Fig. 6 Flow cytometry measurements on the two fluorescent markers anti-BrdU and 7-AAD; (a) Cluster tree with 12 leaves (8 clusters, 4 artifacts); (b) Cluster assignments; core obs have larger labels (c) 90% confidence cluster assignments

5 Summary and Future Work

We have presented a plug-in method for estimating the cluster tree of a density that takes advantage of the ability to exactly compute the level sets of a piecewise constant density estimate. The approach shows flexibility in finding clusters of unequal sizes and shapes. However, the cluster tree is dependent on the (inherently noisy) density estimate. We introduced *clustering with confidence*, an automatic pruning procedure that assesses significance of splits in the cluster tree; the only input needed is the desired confidence level.

These procedures may become computationally intractable as the number of adjacent bins grows with the dimension and are realistically for use in lower dimensions. One high-dimensional approach would be to employ projection or dimension reduction techniques prior to cluster tree estimation. We also have developed a graph-based approach that approximates the cluster tree in high dimensions (Stuetzle and Nugent 2010). CWC then could be applied to the resulting graph to identify significant clusters.

Acknowledgements This work was partially supported by NSF grants DMS-0505824 and DMS-0240019.

References

Buja, A. (2002). Personal communication. Also Buja, A. and Rolke, W. *Calibration for simultaneity: (Re)Sampling methods for simultaneous inference with applications to function estimation and functional data.* In revision.

Cuevas, A., Febrero M., & Fraiman, R. (2000). Estimating the number of clusters. *The Canadian Journal of Statistics, 28*, 367–382.

Cuevas, A., Febrero M., & Fraiman, R. (2001). Cluster analysis: A further approach based on density estimation. *Computational Statistics & Data Analysis, 36*, 441–459.

Fraley, C., & Raftery, A. (1998). How many clusters? which clustering method? Answers via model-based cluster analysis. *The Computer Journal, 41*, 578–588.

Hall, P., & Wand, M. P. (1996) On the accuracy of binned kernel density estimators'. *Journal of Multivariate Analysis, 56*, 165–184.

Hartigan, J. A. (1975). *Clustering Algorithms*. London: Wiley.

Hartigan, J. A. (1981). Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association, 76*, 388–394.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*, 193–218.

Klemelä, J. (2004). Visualization of multivariate density estimates with level set trees. *Journal of Computational and Graphical Statistics, 13*, 599–620.

Lo, K., Brinkman R., & Gottardo, R. (2008). Automated gating of flow cytometry data via robust model-based clustering. *Cytometry, Part A, 73A*, 321–332.

Lo, K., Hahne, F., Brinkman, R.R., and Gottardo, R. (2009). FlowClust: a Bioconductor package for automated gating of flow cytometry data. *BMC Bioinformatics*, *10*, 145.

Mardia, K., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. New York: Academic Press.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*. New York, USA: Marcel Dekker.

Robert, S. (2002). *Algorithms in C, Part 5: Graph Algorithms* (3rd ed.) Reading, MA: Addison-Wesley.

Stuetzle, W. (2003). Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *Journal of Classification, 20*, 25–47.

Stuetzle, W., & Nugent, R. (2010). A generalized single linkage method for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics, 19(2)*.

Walther, G. (1997). Granulometric smoothing. *The Annals of Statistics, 25*, 2273–2299.

Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing*. London: Chapman & Hall.

Wishart, D. (1969). Mode analysis: A generalization of nearest neighbor which reduces chaining effect. In A. J. Cole (Ed.), *Numerical Taxonomy* (pp. 282–311). London: Academic Press.

Local Classification of Discrete Variables by Latent Class Models

Michael Bücker, Gero Szepannek, and Claus Weihs

Abstract "Global" classifiers may fail to distinguish classes adequately in discrimination problems with inhomogeneous groups. Instead, local methods that consider latent subclasses can be adopted in this case. Three different models for local discrimination of categorical variables are presented in this work. They are based on Latent Class Models, which represent discrete finite mixture distributions. Therefore, they can be estimated via the EM algorithm. A corresponding model is constructed analogously to the Mixture Discriminant Analysis by class conditional Latent Class Models. Two other techniques are based on the idea of Common Components Models. Applicable model selection criteria and measures for the classification capability are suggested. In a simulation study, discriminative performance of the methods is compared to that of decision trees and the Naïve Bayes classifier. It turns out that the MDA-type classifier can be seen as a localization of the Naïve Bayes method. Additionally the procedures have been applied to a SNP data set.

1 Introduction

In general, one can not assume homogeneous groups in classification problems. Therefore, one "global" modeling for all classes (as e.g. in Linear Discriminant Analysis) may lead to poor classification results. Hence, in this context models that allow for local structures e.g. through taking account for subclasses should be preferred. An overview on local classification methods can be found in Szepannek et al. (2008). Mixture Discriminant Analysis (MDA) or Common Components (CC) Models are available for continuous variables (see Sect. 2). In Sect. 4 discrete counterparts of these methods will be introduced. The procedures are based on Latent Class Models that are presented in Sect. 3. Model selection criteria are discussed as well as measures for the capability of the Common Components Models. A comparative simulation study shall give an impression of the discriminative power of

M. Bücker (✉)
Fakultät Statistik, Technische Universität Dortmund, 44221 Dortmund, Germany
e-mail: buecker@statistik.tu-dortmund.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_13,

the methods. Subsequently, the techniques are applied to a real life data set of SNP variables (see Sect. 5).

2 Mixtures Versus Common Components

Mixture Discriminant Analysis goes back to Hastie and Tibshirani (1996). Instead of assuming each class to be Gaussian as in Linear Discriminant Analysis the groups are supposed to be a finite mixture of Gaussians. Let the sample space $\Omega = \bigcup_{k=1}^{K} \Omega_k$ be a partition and π_k the prior probability for group k. We aim at predicting the class membership, that is the realization of a random variable $Z \in \{1, \ldots, K\}$, based on the knowledge of the expression of the random vector $X = (X_1, \ldots, X_D)^\top$. We assume X to be a finite mixture of M_k multivariate normals in class k. Thus, the class conditional density is

$$f(x|Z = k) = \sum_{m=1}^{M_k} w_{mk} \phi(x; \mu_{mk}, \Sigma),$$

where $\phi(x; \mu, \Sigma)$ is the pdf of a multivariate Gaussian distribution with mean μ and covariance matrix Σ. The mixture weights w_{mk} and the unknown distribution parameters can be estimated via the EM algorithm.

In the *Common Components Model* the class specific densities are assumed to be

$$f(x|Z = k) = \sum_{m=1}^{M_k} w_{mk} \phi(x; \mu_m, \Sigma_m).$$

Again, the unknown parameters can be estimated via the EM algorithm. The component densities $\phi(x; \mu_m, \Sigma_m)$ do not depend on k but are common components. Only the mixture weight w_{mk} depends on the group.

Plugging in the parameter estimates, the Bayesian decision rule for both models leads to determining the unknown class by

$$\widehat{k}(x) = \arg \max_{k=1,\ldots,K} \frac{\widehat{\pi}_k \widehat{f}(x|Z = k)}{\sum_{l=1}^{K} \widehat{\pi}_l \widehat{f}(x|Z = l)}. \tag{1}$$

3 Latent Class Analysis

Latent Structure Analysis (LSA) was developed by P. F. Lazarsfeld (cf. Lazarsfeld and Henry 1968). *Latent Class Analysis* (LCA) is a special case of LSA where *latent* (unobservable) and *manifest* (observable) variables are discrete. In Hagenaars and McCutcheon (2002) many applications and expansions of the LCA can be found.

For the *Latent Class Model* (LCM) it is assumed that

1. Y is a one dimensional discrete random variable, the latent variable, with support $\{1, \ldots, M\}$ and $P(Y = m) := w_m$ with restriction $\sum_{m=1}^{M} w_m = 1$.
2. The random vector $X = (X_1, \ldots, X_D)^\top$ of manifest variables has a discrete distribution. Any random variable X_d takes values in $\{1, \ldots, R_d\}$. The probability mass function (pmf) of $X_d|Y$ is given by $f(x_d|m) = \prod_{r=1}^{R_d} \theta_{mdr}^{x_{dr}}$, where X_{dr} equals 1 if $X_d = r$ and 0 otherwise; $\theta_{mdr} = P(X_d = r|Y = m)$.
3. The manifest variables $X_1, \ldots, X_D$ are locally independent, that means they are stochastically independent conditioned on Y. So the pmf of $X|Y$ can thus be written as $f(x|m) = \prod_{d=1}^{D} \prod_{r=1}^{R_d} \theta_{mdr}^{x_{dr}}$. It follows directly that the pmf of X is given by

$$f(x) = \sum_{m=1}^{M} w_m f(x|m) = \sum_{m=1}^{M} w_m \prod_{d=1}^{D} \prod_{r=1}^{R_d} \theta_{mdr}^{x_{dr}} \tag{2}$$

and we see that this defines a class of finite mixture distributions.

The assumption of local independence seems to be farfetched and not rational in various situations but this does not cause any problem as any discrete probability distribution can be approximated by the class of discrete finite mixtures given by (2) if M is sufficiently large (cf. Grim and Haindl 2003).

3.1 Estimation

If the number of components M ist not fixed a priori the class of LCMs is not identifiable which can be derived from results in Elmore and Wang (2003) and Teicher (1967). As we see from (2), the LCM is a finite mixture of products of multinomials $\mathbb{M}(1, \theta_{md1}, \ldots, \theta_{mdR_d})$. As established in Elmore and Wang (2003) mixtures of M multinomials $\mathbb{M}(N, \theta_1, \ldots, \theta_p)$ are identifiable if and only if $N \geq 2M - 1$. Combined with a result given by Teicher (1967) that mixtures of the product of marginal distributions are identifiable if and only if mixtures of the marginal distributions are when M is not bounded, we see that the LCM is not identifiable in this case. Nevertheless, when we restrict the number of parameters in advance so that $M\left(\sum_{d=1}^{D} R_d - D + 1\right) - 1 < N$ (no. of parameters < no. of observations) holds, we avoid the problem of non-identifiability.

For LCMs several estimation procedures have been proposed. The most convenient method may be the EM algorithm. The EM steps for estimation of the unknown parameters given the sample $x_1, \ldots, x_N$ are

E step Determination of the conditional expectation of Y given $X = x_n$

$$\tau_{mn} = \frac{w_m f(x_n|m)}{f(x_n)}.$$

M step Maximization of the log-Likelihood and estimation of

$$w_m = \frac{1}{N}\sum_{n=1}^{N}\tau_{mn} \quad \text{and} \quad \theta_{mdr} = \frac{1}{Nw_m}\sum_{n=1}^{N}\tau_{mn}x_{ndr}.$$

3.2 Model Selection

In practice, the number of latent classes is generally unknown and thus also a parameter to identify. With increasing number of latent classes the flexibility but also the complexity, i.e. the number of unknown parameters increases. Hence, there is a necessity of regularization. Among others, the well-known information criteria AIC and BIC can serve for model selection in LCA. Furthermore, goodness-of-fit tests can be applied to examine the conformance of the model and the data. Pearson's χ^2 test statistic or the likelihood ratio χ^2 can be used to compare the fit of different models. In general, distributional assumptions are not met so that instead of calculating p-values one should compare different models and choose the one with the best fit, where overfitting should be considered.

4 Local Classification of Discrete Variables

We can use the former results in order to construct a local discrimination method like the MDA or Common Components Models for categorical data.

4.1 Class Conditional Mixtures

If we assume class conditional mixtures analogously to the MDA and we therefore use LCMs the latent classes represent subclasses. Thus we have as class conditional model

$$P(X = x|Z = k) = f_k(x) = \sum_{m=1}^{M_k} w_{mk}\prod_{d=1}^{D}\prod_{r=1}^{R_d}\theta_{mkdr}^{x_{kdr}},$$

where $x_{kdr} = 1$ if $x_d = r$ in class k and $x_{kdr} = 0$ otherwise; $\theta_{mkdr} = P(X_d = r|Y = m, Z = k)$. We can estimate the model mentioned above by class-wise application of the EM algorithm described in Sect. 3.1. As the method for assigning a class to a new object we choose the one that maximizes its posterior [cf. (1)].

4.2 Common Components

If we take the Common Components Model as a basis for the design of a discrete classifier we would consider the model

$$P(X = x|Z = k) = f_k(x) = \sum_{m=1}^{M} w_{mk} \prod_{d=1}^{D} \prod_{r=1}^{R_d} \theta_{mdr}^{x_{dr}},$$

where only the mixture weight w_{mk} depends on the class membership. In this case the EM procedure needs to be modified since w_{mk} has to be estimated class dependent unlike the parameters θ_{mdr}. Hence the EM steps are

E step Determination of the conditional expectation

$$\tau_{mkn} = \frac{w_{mk} f(x_n|m)}{f(x_n)}.$$

M step Maximization of the log-Likelihood and estimation of

$$w_{mk} = \frac{1}{N_k} \sum_{n=1}^{N_k} \tau_{mkn} \quad \text{and} \quad \theta_{mdr} = \sum_{k=1}^{K} \frac{1}{N_k w_{mk}} \sum_{n=1}^{N_k} \tau_{mkn} x_{ndr}.$$

We call this model CC1. Now let π_k be the prior for class k. Then we have

$$P(X = x) = \sum_{k=1}^{K} \pi_k \sum_{m=1}^{M} w_{mk} \prod_{d=1}^{D} \prod_{r=1}^{R_d} \theta_{mdr}^{x_{dr}}. \tag{3}$$

With $w_{mk} = P(m|k)$ we can define $w_m := P(m) = \sum_{k=1}^{K} P(k)P(m|k) = \sum_{k=1}^{K} \pi_k w_{mk}$. Thus, we can convert (3) as follows

$$P(X = x) = \sum_{m=1}^{M} \sum_{k=1}^{K} \pi_k w_{mk} \prod_{d=1}^{D} \prod_{r=1}^{R_d} \theta_{mdr}^{x_{dr}} = \sum_{m=1}^{M} w_m \prod_{d=1}^{D} \prod_{r=1}^{R_d} \theta_{mdr}^{x_{dr}},$$

which means that we get a global LCM that is group independent. Hence, besides estimating a Common Components Model by the EM steps mentioned above we can also determine a global LCM and estimate the w_{mk} in a second step by $\widehat{w}_{mk} = \frac{1}{N_k} \sum_{i=1}^{N_k} \widehat{P}(Y = m|Z = k, X = x_i)$, with $i = 1, \ldots, N_k$ the index of the observations in group k (cf. Titsias and Likas 2001). This model will be called CC2. Posterior probabilities for CC1 and CC2 can be determined by (1).

4.2.1 Classification Capability

In the Common Components Models subclasses can be compared to end nodes in decision trees. We denote the probability for an object in any subclass m to belong to class k by $w_{km} := P(k|m) = \pi_k w_{mk} / \sum_{k=1}^{K} \pi_k w_{mk}$. The purer the subclasses are, the better the classification result will be. This yields to the idea of applying a measure for the capability of CC Models according to the purity measures used by decision trees like the standardized mean entropy impurity measure $H = -\sum_{m=1}^{M} w_m \sum_{k=1}^{K} w_{km} \cdot \log_K (w_{km})$ with $0 \leq H \leq 1$ or the mean Gini impurity measure $G = \sum_{m=1}^{M} w_m \left[1 - \sum_{k=1}^{K} (w_{km})^2\right]$ with $0 \leq G \leq 1$. Another option is the use of a χ^2 test to verify dependence of subclass membership and class membership which argues for a good classification aptitude of the model.

5 Application

5.1 Simulation Study

First, we examine the properties of the presented classification methods in order to ascertain how different data situations are handled by the methods. For this purpose, we simulate train and test data sets with two classes with different properties as mentioned below. We choose a full factorial design with 10 repetitions of each trial and analyse the significant effects on the misclassification rate (estimated on a test data set) by an ANOVA. The dependent variable is chosen to be a transformation of the estimated classification rate $1 - \widehat{\varepsilon}$, namely the log-odds $\log\left(\frac{1-\widehat{\varepsilon}}{\widehat{\varepsilon}}\right)$ so that it ranges over $\mathbb{R}$. The results are compared to the theoretical Bayes error rate and to the performance of a Naïve Bayes classifier and classification trees. The two data generating processes (MDA-type model and CC) must be examined separately:

Class-wise LCMs: For class-wise LCMs the data is generated as follows:

1. Choose a class $k \in \{1, \ldots, K\}$ with probability π_k.
2. Choose a subclass $m \in \{1, \ldots, M_k\}$ with probability w_{mk}.
3. Choose $r \in \{1, \ldots, R_d\}$ for $X_d \in \{X_1, \ldots, X_D\}$ with probability θ_{mkdr}.

The estimated effects on the classification rate relate to class priors (different or equal in each class), subclass priors (different or equal in each subclass), number of observations (500 or 5,000), overlap of the classes (controlled by the probabilities θ_{mkdr}) and the number of irrelevant variables (variables that do not differ in each class).

The results reveal that only class overlap and the number of irrelevant variables have significant effects. The more overlapping the subclasses are and the more irrelevant variables occur the higher the misclassification rate in all investigated classification methods becomes. The mean misclassification rates and the Bayes

error rate for all 320 trials are shown in Table 1. We see that the discrete MDA has the lowest misclassification rate and that the first Common Components approach shows a very high error rate.

CC Models: For the CC Models the data generating steps are:

1. Choose a latent class $m \in \{1, \ldots, M\}$ with probability w_m.
2. Choose independently a class $k \in \{1, \ldots, K\}$ with probability w_{km} and for $X_d \in \{X_1, \ldots, X_D\}$ a value $r \in \{1, \ldots, R_d\}$ with probability θ_{mdr}.

The estimated effects on the classification rate relate to class priors (different or equal in each class), number of observations (500 or 5,000), overlap of the classes (controlled by the probabilities θ_{mdr}), the number of irrelevant variables and the subclass purity (controlled by w_{km}).

The results reveal that class overlap, the number of irrelevant variables and the subclass purity have significant effects. The more overlapping the classes are, the more irrelevant variables occur and the impurer the subclasses are the higher the misclassification rate in all investigated classification methods becomes. The results in Table 1 reveal that the discrete MDA and the second Common Components approach show low error rates. In both simulations the discrete MDA misclassification level is close to the Bayes error rate. The first Common Components approach classifies inadequately in the data situation simulated by class-wise LCMs but is acceptable in the CC data situation.

Discrete MDA as localization of Naïve Bayes: The fact that the discrete MDA assumes local independence while Naïve Bayes supposes that the variables are independent per group (i.e. "global independence") suggests that the discrete MDA could be seen as a localization of Naïve Bayes. We investigate this by a simple simulation. We generate data sets consisting of two classes and two subclasses based on two variables with four outcomes each. However, we choose the parameters θ_{mkdr} in a way that the subclasses do not really differ. Indeed, we can combine two outcomes of the variables respectively and the subclasses will disappear since the probabilities for these combinations are the same in each subclass (i.e. $\theta_{1kd1} + \theta_{1kd2} = \theta_{2kd1} + \theta_{2kd2}$ and $\theta_{1kd3} + \theta_{1kd4} = \theta_{2kd3} + \theta_{2kd4}, k, d = 1, 2$). The Bayes error for this situation is 0.2. The discrete MDA has a error rate of 0.2058 and Naïve Bayes of 0.2044. In situations of existing subclasses we discovered quite different error rates for these methods.

Table 1 Mean error rates for data generated by the MDA-type model and the Common Components Model

Data generated by	Bayes error	Discrete MDA	CC1	CC2	Naïve Bayes	CART
Class-wise LCM	0.123	0.148	0.222	0.160	0.195	0.164
CC	0.254	0.264	0.280	0.267	0.259	0.271

5.2 SNP Data

In a next step the presented procedures are applied to a SNP data set to investigate their practicability in real data situations. The analyzed data originates from the GENICA study (cf. Brauch et al. 2000), which is an age-matched, population-based, case-control candidate SNP study. It aims at identifying genetic and gene-environment associated breast cancer risks. The data contains 1,166 observations, 605 controls and 561 cases, of 68 SNP variables and six categorical epidemiological variables. We compare our classification results with the results mentioned in Schiffner et al. (2009). Therefore, the same partition of the data set for estimating the error rates by cross-validation is used. For computational reasons we restrict the number of subclasses per group to a maximum of 10.

The misclassification rate of the discrete MDA (**0.220**, standard deviation 0.030) out-performs the best rate of Schiffner et al. (2009) (logistic regression: **0.366**). CC1 (**0.471**, sd 0.049) and CC2 (**0.345**, sd 0.056) have significantly higher error rates. We therefore conclude that the discrete MDA appears to be an adequate model to classify the SNP data. The CC methods seem to be less appropriate in this case. A reason for this fact could be the impurity of the subclasses ($H = 0.99$ and $G = 0.50$) which might entail an inaptitude of these methods.

6 Conclusion

We presented three models based on LCMs that provide a flexible approach to local classification. The assumption of local independence allows for relatively sparse data sets in contrast to using mixtures of multinomials. The models can handle missing values without imputation. The discrete MDA can be seen as a localized version of the Naïve Bayes method.

Further efforts could extend the methods to mixed data assuming normality of the continuous variables. A clustering technique based on this idea can be found in Hunt and Jorgensen (2003) which could be used for local classification of mixed data.

References

Brauch, H., Brüning, Th., Hamann, U. & Ko, Y. (2000). GENICA Network. URL http://www.genica.de.

Elmore, R., & Wang, S. (2003). *Identifiability and estimation in finite mixture models with multinomial components*. Technical Report 03–04, Department of Statistics, Pennsylvania State University.

Grim, J., & Haindl, M. (2003). Texture modelling by discrete distribution mixtures. *Computational Statistics & Data Analysis, 41*, 603–615.

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge: Cambridge University Press.

Hastie, T., & Tibshirani, R. (1996). Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society B, 58*, 155–176.

Hunt, L., & Jorgensen, M. (2003). Mixture model clustering for mixed data with missing information. *Computational Statistics & Data Analysis, 41*, 429–440.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Miflin.

Schiffner, J., Szepannek, G., Monthé, Th., & Weihs, C. (2009). Localized logistic regression for categorical influential factors. To appear in A. Fink, B. Lausen, W. Seidel and A. Ultsch (Eds.), *Advances in data analysis, data handling and business intelligence*. Heidelberg-Berlin: Springer-Verlag.

Szepannek, G., Schiffner, J., Wilson, J., & Weihs, C. (2008). Local modelling in classification. In P. Perner (Ed.), *Advances in data mining, Lecture Notes in Artificial Intelligence* (Vol. 5077, pp. 153–164). Heidelberg: Springer-Verlag.

Teicher, H. (1967). Identifiability of mixtures of product measures. *The Annals of Mathematical Statistics, 38*, 1300–1302.

Titsias, M. K., & Likas, A. C. (2001). Shared kernel models for class conditional density estimation. *IEEE Transactions on Neural Networks, 12*, 987–997.

A Comparative Study on Discrete Discriminant Analysis through a Hierarchical Coupling Approach

Ana Sousa Ferreira

Abstract In discrete discriminant analysis the high-dimensionality problem often causes discriminant methods to perform poorly, specially in the multiclass case. The Hierarchical Coupling Model (HIERM) enables a reduction of the multiclass problem into several biclass problems embedded in a binary tree. With this approach, at each level of the tree, the basic affinity coefficient is used to select the new couple of classes among all possible forms of merging the original classes. After identifying the pair of classes to be considered, the decision rule for this biclass problem is based on the combining model that minimizes the error rate. The performance of this model leads to a considerable improvement in the misclassification error rate. Furthermore, its representation is appealing which makes it easily interpretable. In this study we propose to explore the comparison of the HIERM model with other models where the choice of the decomposition at each level of the binary tree among all possible forms of merging is made using one of the traditional similarity coefficients in Cluster Analysis. The performance of HIERM and the new models is compared on real data.

1 Introduction

Let $X = (x_1, \ldots, x_n)$ denotes a n-dimensional training sample of multivariate observations, associated with p discrete variables, for which each object is assumed to become from one of K exclusive classes $G_1, G_2, \ldots, G_K$ with prior probabilities $\pi_1, \pi_2, \ldots, \pi_K$ ($\sum_{k=1}^{K} \pi_k = 1$).

The context of this study is Discrete Discriminant Analysis (DDA) and we will be focusing on the small sample setting and binary variables.

A.S. Ferreira
LEAD, FPCE, University of Lisbon, Alameda da Universidade, 1649-013 Lisboa,
CEAUL, Multivariate Data Analysis and Modelling Project
e-mail: asferreira@fpce.ul.pt

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_14,

We consider a classical DDA problem: our main aim is to derive a classification rule to allocate subjects, whose class is unknown, to one of the K classes, whose members are described by the p binary variables. The Bayes classification rule assigns an individual vector $\underline{x}$ into G_k if

$$\pi_k P(\underline{x}|G_k) \geq \pi_l P(\underline{x}|G_l) \; for \; l = 1, \ldots, K \; and \; l \neq k \tag{1}$$

where $P(\underline{x}|G_l)$ denotes the conditional probability function for the $l - th$ class. Usually, the conditional probability functions are unknown and are estimated on the basis of the training sample.

For discrete problems the most natural model is to assume that the class conditional functions $P(\underline{x}|G_l)$ are multinomial probabilities estimated by the observed frequencies in the training set. However, model FMM involves $2^p - 1$ parameters in each class. Hence, even for moderate p, not all of the parameters are identifiable. One way of dealing with this high-dimensionality problem consists of reducing the number of parameters to be estimated. The FOIM model assumes that the p binary variables are independent in each class G_k, $k = 1, \ldots, K$. So, the number of parameters to be estimated for each class is reduced from $2^p - 1$ to p. This method is simple but may be unrealistic in many situations.

In Discrete Discriminant Analysis there is a high-dimensional problem due to the large number of parameters to be estimated in most of the models. Furthermore, if we have small or moderate sample sizes, we encounter a problem of sparseness in which some of the multinomial cells may have no data in the training sets. Thus, most of the discrete discrimination methods perform poorly. This high-dimensional problem is even more complex in the multiclass case.

2 Combining Models in Biclass Problems

In many situations we have several classification rules in competition for the same problem and one of those rules is selected, based on some criteria. Acting in such a way leads to rejecting several classification rules for which the parameters have been estimated. Besides, misclassified subjects can be different for the different models. Thus, those rules may contain useful information on the classification problem, and this information is lost by selecting a single model. The idea of combining models is currently emerging in an increasing number of papers, with a view to obtaining a more robust and stable model than any of the competing models.

In biclass problems we proposed (Sousa Ferreira 2000) a classification rule based on a combining model: an intuitive combination method is to propose a single coefficient, producing an intermediate model between the FMM model and the FOIM model:

$$\hat{P}_k(\underline{x}|\beta) = \beta \hat{P}_{FOIM}(\underline{x}|G_k) + (1 - \beta)\hat{P}_{FMM}(\underline{x}|G_k) \tag{2}$$

We proposed and evaluated several strategies to estimate the coefficient β (e.g., Sousa Ferreira 2000; Brito et al. 2006). A natural way of deriving the coefficient β is by minimizing the fitting error using a least squares criterion. The Committee of Methods introduced by (Bishop 1995) in the neural computing literature is such an approach. Another strategy is a measure of the relative performance of the FOIM model that takes account of model uncertainty (Raftery 1996) and is based on the integrated likelihoods for the FOIM and FMM models. In this study we will use an approach to estimate the β coefficient using a least squares regression criterion (Leblanc and Tibshirani 1996):

$$\widehat{\beta}_{LSR} = \frac{\sum_{i=1}^{n}(C_2(x_i) - C_1(x_i))C_2(x_i) - \sum_{i=1}^{n} y_i(C_2(x_i) - C_1(x_i))}{\sum_{i=1}^{n}(C_2(x_i) - C_1(x_i))^2} \tag{3}$$

where C_1, C_2 represent, respectively, the *a posteriori* probabilities for FOIM and FMM models estimated by cross-validation. The combining models approach using a single coefficient proved to be a good alternative for reducing the dimensionality problem in the two classes case.

3 The Hierarchical Coupling Model (HIERM)

We proposed a method, inspired by Friedman's approach (Friedman 1996), for reducing the multiclass problem into several biclass problems (Sousa Ferreira et al. 1999) embedded in a binary tree. HIERM needs two decisions at each level:

- Selection of the hierarchical coupling among the $2^{K-1} - 1$ possible class couple;
- In each node of the tree, choice of the combining model that gives the best classification rule for the chosen couple.

The individual vector $\underline{x}$ is assigned to the class associated with the last node of the tree on which $\underline{x}$ falls. The main aim of this approach is to obtain a better prediction performance and more results stability.

At the beginning we have K training subsamples and we want to reorganize these K classes into two classes. So, we propose either to explore all the hierarchical coupling solutions or to select the two new classes that are the most separable.

The basic affinity coefficient (Matusita 1955; Bacelar-Nicolau 1985) can be used to choose the hierarchical coupling at each level of the tree, $F_1 = \sqrt{p_j}$ and $F_2 = \sqrt{q_j}$, $j = 1, 2, \ldots, p$ being two discrete distributions defined in the same space:

$$\rho(F_1, F_2) = \sum_{l=1}^{L} \sqrt{p_j}\sqrt{q_j} \tag{4}$$

and is easily computed in our classification problem.

Table 1 Frequencies for each pair of subjects, across all variables

		Subject *i*	
		1	0
	1	a	b
Subject *j*			
	0	c	d

4 Comparison of the HIERM Model with Other Models, Using Similarity Coefficients for Binary Data

In the previously defined HIERM model the basic affinity coefficient is used to select the new couple in each level of the binary tree. However, since there are several similarity coefficients for binary data, this choice can also be made using one of them (Hubálek 1982). The aim of this study is to explore the comparison of the HIERM model with other models where the choice of the couple at each level of the binary tree is made using one of the traditional similarity coefficients for binary data in Cluster Analysis.

4.1 Similarity Coefficients for Binary Data

As we know, in order to analyze subject or variable similarities we use their descriptions by p independent variables or n subjects. In the case of binary data, we can summarize the information for each pair of subjects, each taken from one of the classes of the chosen couple (G_l, G_k), on a 2×2 contingency table (see Table 1), where a, b, c, d are frequencies, across all variables, respectively positive co-occurrences, occurrence/non-occurrence, non-occurrence/occurrence and negative co-occurrences.

In order to explore the performance of the HIERM model and the new models using traditional similarity coefficients for binary data, we selected ten similarity coefficients.

We call *Type I Coefficients* the five similarity coefficients selected, that exclude the negative co-occurrences, defined in Table 2 and *Type II Coefficients* the five similarity coefficients selected, that include the negative co-occurrences, defined in Table 3.

5 Numerical Experiments

The performance of HIERM and the new models using the ten similarity coefficients has been compared on both real and simulated binary data. However, for the sake of simplicity and dimension of this study, we only present here, the application to real data.

Table 2 Type I similarity coefficients

Type I Coefficients	Definition	Occurrence Interval
Jaccard (1901)	$\frac{a}{a+b+c}$	[0,1]
Kulczynski I (1927)	$\frac{a}{b+c}$	[0,+∞[
Kulczynski II (1927)	$\frac{\frac{a}{2}(2a+b+c)}{(a+b)(a+c)}$	[0,1]
Dice and Sorensen (1945,1948)	$\frac{2a}{2a+b+c}$	[0,1]
Sokal and Sneath II (1963)	$\frac{a}{a+2(b+c)}$	[0,1]

Table 3 Type II similarity coefficients

Type II Coefficients	Definition	Occurrence Interval
Simple Matching (1958)	$\frac{a+d}{a+b+c+d}$	[0,1]
Russel and Rao (1940)	$\frac{a}{a+b+c+d}$	[0,1]
Rogers (1960)	$\frac{a+d}{a+d+2(b+c)}$	[0,1]
Haman (1961)	$\frac{(a+d)-(b+c)}{a+b+c+d}$	[−1,1]
Sokal and Sneath I (1963)	$\frac{2(a+d)}{2(a+d)+b+c}$	[0,1]

We thank our colleagues for providing data GSS1 ("GSS1" is a registered trademark authorized for Portuguese adaptation by R. Pires, FPCE, University of Lisbon), ALEXITH (Prazeres 1998) and CAREER (Lima 1998).

Since HIERM was proposed on the small or moderate sample size settings, the proposed models' performance is evaluated by two-fold cross validation or the error rate is estimated in a test sample.

In this study, the *a priori* class probabilities are taken to be equal. Characterization of data sets are summarize in Table 4.

As was previously mentioned, the affinity coefficient is easily computed in our kind of data.

However, using the ten similarity coefficients in order to choose the new couple at each level of the tree, we need one aggregation criteria to obtain the measure of similarity between each pair of classes. We selected the Average Linkage Criteria since it is one of the most commonly used that best achieves our objectives. This criteria is computed as the average similarities between subjects from the first class and subjects from the second. The averaging is performed over all pairs of subjects. In the context of this study, for instance Single Linkage and Complete Linkage leaves too often towards the extreme values zero or one.

Table 5 summarizes the results of the choice of the hierarchical coupling by *Type I coefficients* and Table 6 describes the same results for *Type II coefficients*.

Table 4 Characterization of data sets

Data sets	Description	n	Groups	Variables
GSS1	Gudjonsson suggestibility scales three age groups	98	$n_1 = 72$ $n_2 = 7$ $n_3 = 19$	6 variables
ALEXITH	Clinical psychological data three groups of Alexithymia degrees	34	$n_1 = 14$ $n_2 = 13$ $n_3 = 7$	6 variables
CAREER	Psychological counselling career data four degree courses	600	$n_1 = 119$ $n_2 = 212$ $n_3 = 148$ $n_4 = 121$	10 variables

Table 5 Choice of the hierarchical coupling by Type I coefficients

Data sets	***Type I coefficients***					
	Affinity	Jaccard	Kul. I	Kul. II	Dice S.	Sokal II
GSS1	.4056	.1595	.2718	.2590	.2198	.1085
	.2613	**.0788**	**.1249**	**.1343**	**.1101**	**.0524**
	.3529	.1898	.3286	.3111	.2628	.1283
ALEXITH	.4358	**.3296**	.6268	.5000	**.4265**	**.2399**
	.6456	.3720	.6624	.5488	.5029	.3077
	.4346	.4022	**.5723**	**.4999**	.4647	.2861
CAREER			1^{st} **level of the tree**			
	.9148	**.1355**	**.1420**	**.1785**	**.1620**	**.1112**
	.9240	.1533	.1873	.2409	.2172	.1509
	.9528	.1948	.2150	.2601	.2347	.1583
	.9195	.1686	.1631	.2201	.1999	.1401
	.9636	.1553	.1705	.2068	.1869	.1262
	.9398	.1820	.1941	.2408	.2178	.1490
	.9389	.1584	.1727	.2092	.1898	.1294
			2^{nd} **level of the tree**			
	.9784	.1463	.1408	.1909	.1729	.1220
	.9689	.1463	.1526	.1934	.1746	.1203
	.9534	**.1270**	**.1279**	**.1671**	**.1511**	**.1049**

It may be noted that in the case of K = 3 classes *a priori* there are 3 hierarchical coupling solutions, but in the case of K = 4 classes *a priori* there are 7.

The results of Table 5 show that, for almost all data, coefficients of Type I - Jaccard, Kulczynski I, Kulczynski II, Dice-Sorensen and Sokal and Sneath I - choose the same hierarchical coupling solution as the basic affinity coefficient, probably due to the fact that all of them excluded negative co-occurences.

In Table 6 similar selections are also observed in Simple Matching, Rogers, Haman and Sokal and Sneath I, probably due to the fact that they included the negative co-occurences.

Table 6 Choice of the hierarchical coupling by Type II coefficients

Data sets	*Type II coefficients*					
	Affinity	Jaccard	Kul. I	Kul. II	Dice S.	Sokal II
GSS1	.4056	.5671	.0939	.4280	.1343	.6981
	.2613	.6243	**.0374**	.4851	.2486	.7455
	.3529	**.5339**	.1144	**.3954**	**.0677**	**.6687**
ALEXITH	.4358	**.5488**	.1577	**.4047**	**.0976**	**.6880**
	.6456	.5842	.1734	.4428	.1685	.7173
	.4346	.5864	**.1279**	.4443	.1728	.7173
CAREER			1^{st} **level of the tree**			
	.9148	.6641	**.0680**	.5569	.3282	.7613
	.9240	.6657	.0893	.5571	.3315	.7628
	.9528	.6347	.1015	.5216	.2694	.7382
	.9195	.6544	.0811	.5434	.3089	.7546
	.9636	.6045	.0790	.5032	.2420	.6958
	.9398	.6665	.0896	.5564	.3330	.7647
	.9389	**.5568**	.0796	**.4618**	**.1603**	**.6423**
			2^{nd} **level of the tree**			
	.9784	.6845	.0698	.5799	.3691	.7773
	.9689	**.6780**	.0728	**.5733**	**.3561**	**.7714**
	.9534	.6793	**.0627**	.5752	.3587	.7724

Table 7 Performance comparison of Type I and Type II Hierarchical coefficient models for GSS1 data

1^{st} level	2^{nd} level		FOIM	FMM	KER	HIERM	HIERM
Type I Coeff.	$G_1 vs. G_3$	Error rate	.38	.59	.38	.21	**.21**
G_2 **vs.** G_1**+**G_3	λ			.95	1.000	.95	
	β - 1^{st}				.3349	.3400	
	- 2^{nd}				1.000	1.000	
Type II Coeff.	G_1 **vs.**G_2	Error rate	.38	.59	.38	.17	**.14**
G_3 **vs.**G_1**+**G_2	λ			.95	1.000	.95	
	β - 1^{st}				.5877	.5881	

The Russel and Rao coefficient presents different choices to them all, the coefficients of type II, probably due to the fact that it excluded the negative co-occurences in the numerator and included it in the denominator of its expression.

Thus, almost always, two types of hierarchical coupling solutions are achieved by these eleven coefficients: Affinity, Type I and Russel and Rao choose one and Type II coefficients select another one.

Therefore, after choosing the hierarchical coupling in each level of the tree, we compare the performance of the HIERM model using the two types of hierarchical coupling solutions, for the three real data sets in Tables 7, 8 and 9.

In Table 7, for GSS1 data, the choice of Type II coefficients leads to the miminum error.

Table 8 Performance comparison of Type I and Type II Hierarchical coefficient models for ALEXITH data

1^{st} level	2^{nd} level		FOIM	FMM	KER	HIERM	HIERM
Type I Coeff.	$G_1 vs. G_2$	Error rate	.53	.70	.56	**.32**	.35
G_3 **vs.** G_1+G_2	λ			.95	1.000	.95	
	β - 1^{st}				.1800	.9902	
	- 2^{nd}				.4401	.9809	
Type II Coeff.	G_2 **vs.**G_3	Error rate	.53	.70	.56	**.47**	.47
G_1 **vs.**G_2+G_3	λ			.95	1.000	.95	
	β - 1^{st}				1.000	.9100	
	- 2^{nd}				.8100	.0900	

Table 9 Performance comparison of Type I and Type II Hierarchical coefficient models for CAREER data

1^{st} level	2^{nd} level	3^{rd} level	FOIM	FMM	KER	HIERM	HIERM	
Type I Coeff.	G_3**vs.**G_1+G_2	G_1**vs.**G_2	Error rate	.66	.67	.65	**.10**	.25
G_1 **vs.** G_2+G_3+G_4	λ			.95	1.000	.95		
	β - 1^{st}				.9171	.0000		
	- 2^{nd}				1.000	.0000		
	- 3^{rd}				.9547	.0000		
	G_2**vs.**G_1+G_3	G_1 **vs.**G_2	Error rate	.66	.67	.65	.53	.34
	β - 1^{st}				.9171	.0000		
	- 2^{nd}				.7606	.0000		
	- 3^{rd}				.8883	.0000		
G_2+G_3 **vs.**G_1+G_4	λ			.95	1.000	.95		
	β - 1^{st}				.8139	.9547		
	- 3^{rd}				.8892	.0000		

In Tables 8 and 9 for ALEXITH and CAREER data, the choice of Type I coefficients leads to the miminum error.

6 Conclusions

The comparison made showed that using the affinity coefficient or Type I similarity coefficients to select the hierarchical coupling in each branch of the tree is a good option, since they reveal good performance in the small or moderate sample setting. These models frequently provide the lowest estimates of the misclassification risk, even in simulated data.

Due to its easy application in the HIERM model, the basic affinity coefficient takes advantage.

In the multiclass case, the HIERM approach leads to a considerable improvement of the misclassification error rate. Furthermore its representation is appealing which makes it easily interpretable. Due to the fact that Type I coefficients lead, for almost all data, to decompositions with the best misclassification error rates, we suggest using one of them or the basic affinity coefficient in the HIERM model.

References

Bacelar-Nicolau, H. (1985). The affinity coefficient in cluster analysis. *Mathematics of Operations Research, 53*, 507–512.

Bishop, C. (1995). *Neural networks for pattern recognition*. London: Oxford University Press.

Brito, I., Celeux, G., & Sousa Ferreira, A. (2006). Combining methods in supervised classification: a comparative study on discrete and continuous problems. *REVSTAT - Statistical Journal, 4*(3), 201–225.

Friedman, J. F. (1996). *Another approach to polychotomous classification*. Technical Report, Stanford University.

Hubálek, Z. (1982). Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biological Reviews of the Cambridge Philosophical Society, 57*, 669–689.

Leblanc, M., & Tibshirani, R. (1996). Combining estimates in regression and classification. *Journal of the American Statistical Association, 91*, 1641–1650.

Lima, M. R. (1998). *Orientação e Desenvolvimento da Carreira em Estudantes Universitários*. PhD Thesis (in Portuguese), University of Lisbon.

Matusita, K. (1955). Decision rules based on distance for problems of fit, two samples and estimation. *Annals of the Institute of Statistical Mathematics, 26*(4), 631–640.

Prazeres, N. L. (1998). *Ensaio de um Estudo sobre Alexitimia com o Rorschach e a Escala de Alexitimia de Toronto*. Master Thesis (in Portuguese), University of Lisbon.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika, 83*, 251–266.

Sousa Ferreira, A. (2000). *Combining models in discrete discriminant analysis*. PhD Thesis (in Portuguese), New University of Lisbon.

Sousa Ferreira, A., Celeux, G., & Bacelar-Nicolau, H. (1999). Combining models in discrete discriminant analysis by a hierarchical coupling approach. In H. Bacelar-Nicolau, F. Costa Nicolau & J. Janssen (Eds.), *ASMDA 99* (pp. 159–164). Lisboa: INE.

A Comparative Study of Several Parametric and Semiparametric Approaches for Time Series Classification

Sonia Pértega Díaz and José A. Vilar

Abstract Several non-parametric statistics originally designed to test the equality of the log-spectra of two stochastic processes are proposed as dissimilarity measures between two time series. Their behavior in time series clustering is analyzed throughout a simulation study, and compared with the performance of several model-free and model-based dissimilarity measures. Up to three different classification settings are considered: (1) to distinguish between stationary and non-stationary time series, (2) to classify different ARMA processes and (3) to classify several non-linear time series models. As it was expected, the performance of a particular dissimilarity metric strongly depended on the type of processes subjected to clustering. Among all the measures studied, the non-parametric distances showed the most robust behaviour.

1 Introduction

Time series clustering is an important area of research with applications in many fields, see e.g. the classification of industrial production series (Piccolo 1990), the comparison of seismological data (Kakizawa et al. 1998) or clustering of banks according their weekly share price (Vilar et al. 2009).

A key point in cluster analysis is to determine a distance measure between objects, so the existing clustering algorithms can be directly used. However, the high dimensionality and the underlying autocorrelation structure are features of time dependent data that become more complicated cluster analysis.

Previous arguments account for the increase in the number of studies on time series clustering in the last years (Caiado et al. 2006; Kakizawa et al. 1998; Maharaj 1996, 2000; Piccolo 1990; Vilar and Pértega 2004). An excellent overview on time series clustering can be found in Liao (2005).

S.P. Díaz (✉)
Unidad de Epidemiología Clínica y Bioestadística, Hospital de A Coruña, As Xubias, 84, E-15006 A Coruña, Spain
e-mail: Sonia.Pertega.Diaz@sergas.es

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_15,

The present work has two objectives. First, to analyze the behavior in time series clustering of two non-parametric statistics originally designed to test the equality of the log-spectra of two stochastic processes. Second, to extend the comparative analysis performed by Caiado et al. (2006) including new dissimilarity measures and considering the classification of different kinds of processes.

2 Some Dissimilarity Measures Between Time Series

Given $\boldsymbol{X}_T = (X_1, \ldots, X_T)^t$ and $\boldsymbol{Y}_T = (Y_1, \ldots, Y_T)^t$ partial realizations from two scalar-valued processes $X = \{X_t, t \in \mathbb{Z}\}$ and $Y = \{Y_t, t \in \mathbb{Z}\}$, the following measures of dissimilarity between X and Y will be compared along this work:

1. The Euclidean distance, $d_E(X, Y) = \left\{\sum_{t=1}^{T} (X_t - Y_t)^2\right\}^{1/2}$.
2. Two measures based on the estimated autocorrelation functions (ACFs) (see Galeano and Peña 2000). First, the Euclidean distance between the first L estimated autocorrelation coefficients, $d_{ACFU}(X, Y) = \left\{\sum_{i=1}^{L} (\widehat{\rho}_{i,X} - \widehat{\rho}_{i,Y})^2\right\}^{1/2}$. The second ACF metric is defined by introducing geometric weights decaying with the lag, namely $d_{ACFG}(X, Y) = \left\{\sum_{i=1}^{L} p(1-p)^i (\widehat{\rho}_{i,X} - \widehat{\rho}_{i,Y})^2\right\}^{1/2}$, with $p = 0.5$.
3. In a similar way, two distances based on the estimated partial autocorrelation functions (PACF's) (see Caiado et al. 2006): the Euclidean distance between the sample partial autocorrelation coefficients with uniform weights (d_{PACFU}) and with geometric weights decaying with the lag (d_{PACFG}).
4. The metric introduced by Piccolo in Piccolo (1990) for invertible ARIMA processes, defined as the Euclidean distance between their autoregressive expansions. In practice, automatic modeling of AR structures is performed by means of Akaike's Information Criterion (AIC). Thus, the distance is calculated as:

$$d_{PIC}(X, Y) = \left\{ \sum_{j=1}^{max(k_1,k_2)} \left(\widehat{\pi}_{j,X} - \widehat{\pi}_{j,Y}\right)^2 \right\}^{1/2}, \tag{1}$$

with $\widehat{\boldsymbol{\Pi}}_X = \left(\widehat{\pi}_{1,X}, \ldots, \widehat{\pi}_{k_1,X}\right)^t$ and $\widehat{\boldsymbol{\Pi}}_Y = \left(\widehat{\pi}_{1,Y}, \ldots, \widehat{\pi}_{k_2,Y}\right)^t$ the vectors of AR(k_1) and AR(k_2) parameter estimations of the series $\boldsymbol{X}_T$ and $\boldsymbol{Y}_T$.
5. Maharaj's distance for ARMA processes (Maharaj 1996), that is based on a test to determine if two time series have significantly different generating processes:

$$d_M(X, Y) = \sqrt{T} \left(\widehat{\boldsymbol{\Pi}}_X - \widehat{\boldsymbol{\Pi}}_Y\right)^t \widehat{\boldsymbol{V}}^{-1} \left(\widehat{\boldsymbol{\Pi}}_X - \widehat{\boldsymbol{\Pi}}_Y\right), \tag{2}$$

with $\widehat{\boldsymbol{V}}$ an estimator of $\boldsymbol{V} = \sigma_X^2 \boldsymbol{R}_X^{-1}(k) + \sigma_Y^2 \boldsymbol{R}_Y^{-1}(k)$, where σ_X^2 and σ_Y^2 denote the variances of the white noise processes associated with $\boldsymbol{X}_T$ and $\boldsymbol{Y}_T$, and $\boldsymbol{R}_X$ and $\boldsymbol{R}_Y$ are the sample covariance matrices of both series.

6. Distances based on the periodogram. In particular, the Euclidean distance between: the periodogram ordinates ($d_P(X, Y)$), the normalized periodogram ordinates ($d_{NP}(X, Y)$), the logarithm of the periodogram ordinates ($d_{LP}(X, Y)$) and the logarithm of the normalized periodogram ordinates ($d_{LNP}(X, Y)$).
7. The spectral disparity measure d_W defined in Vilar and Pértega (2004) as:

$$d_W(X, Y) = \frac{1}{4\pi} \int_{-\pi}^{\pi} W\left(\frac{\widehat{f}_X(\lambda)}{\widehat{f}_Y(\lambda)}\right) d\lambda, \tag{3}$$

where $W(x) = \log(0.5x + 0.5) - 0.5\log(x)$. Different versions of d_W are considered according to the spectral estimators $\widehat{f}_X$ and $\widehat{f}_Y$ (see Fan and Kreutzberger 1998, for details):

- $d_{W(DLS)}$, when $\widehat{f}_X$ and $\widehat{f}_Y$ are local linear smoothers of the periodograms obtained via least squares;
- $d_{W(LS)}$, when $\widehat{f}_X$ and $\widehat{f}_Y$ are the exponential transformation of the local linear smoothers of the log-periodograms, obtained via least squares;
- $d_{W(LK)}$, when $\widehat{f}_X$ and $\widehat{f}_Y$ are the exponential transformation of the local linear smoothers of the log-periodograms, now obtained using the maximum local likelihood criterion instead of the least squares one.

In all cases, the Epanechnikov kernel was used and the bandwidth determined by cross-validation. In order to obtain a symmetrized version of d_W, we take as divergence function $\tilde{W}(x) = W(x) + W\left(x^{-1}\right)$.

8. Two new measures based on non-parametric statistics originally designed to test the equality of the corresponding log-spectra, m_X and m_Y. First, we focused on a modification of the generalized likelihood ratio test introduced in Fan and Zhang (2004):

$$\begin{aligned} d_{GLK}(X, Y) = &\sum_{k=1}^{n} \left[Z_k - \widehat{\mu}(\lambda_k) - 2\log\left(1 + e^{\{Z_k - \widehat{\mu}(\lambda_k)\}}\right)\right] \\ &- \sum_{k=1}^{n} \left[Z_k - 2\log\left(1 + e^{Z_k}\right)\right], \end{aligned} \tag{4}$$

where $Z_k = \log(I_x(\lambda_k)) - \log(I_y(\lambda_k))$, $\mu(\lambda_k) = m_X(\lambda_k) - m_Y(\lambda_k)$ and $\widehat{\mu}(\lambda_k)$ is the local maximum log-likelihood estimator of $\mu(\lambda_k)$ computed by local linear fitting.

Second, a test statistic based on the Cramér-von-Mises-type functional distance between the estimators of the log-spectra. In particular, we consider

$$d_{CM}(X, Y) = \int (\widehat{m}_X(\lambda) - \widehat{m}_Y(\lambda))^2 \, d\lambda, \tag{5}$$

with $\widehat{m}_X(\lambda)$ and $\widehat{m}_Y(\lambda)$ the local linear smoothers of the log-periodograms, obtained using the maximum local likelihood criterion.

3 Simulation Study

In this section, we present the results from the numerical study designed to compare the behaviour of the measures in Sect. 2 under different classification setups. Three classification problems were considered: (1) to distinguish between stationary and non-stationary time series, (2) to classify different ARMA processes and (3) to classify several non-linear time series models.

3.1 Classification of Time Series as Stationary or Non-Stationary

The first set of experiments was aimed at extending a previous study (Caiado et al. 2006) by including the non-parametric dissimilarity measures proposed in Sect. 2.

As in Caiado et al. (2006), $s = 1$ realization from the following 12 models was generated:

(a) AR(1)	$\phi_1 = 0.9$	(g) ARIMA(1, 1, 0)	$\phi_1 = -0.1$
(b) AR(2)	$\phi_1 = 0.95, \phi_2 = -0.1$	(h) ARIMA(0, 1, 0)	
(c) ARMA(1, 1)	$\phi_1 = 0.95, \theta_1 = 0.1$	(i) ARIMA(0, 1, 1)	$\theta_1 = 0.1$
(d) ARMA(1, 1)	$\phi_1 = -0.1, \theta_1 = -0.95$	(j) ARIMA(0, 1, 1)	$\theta_1 = -0.1$
(e) MA(1)	$\theta_1 = -0.9$	(k) ARIMA(1, 1, 1)	$\phi_1 = 0.1, \theta_1 = -0.1$
(f) MA(2)	$\theta_1 = -0.95, \theta_2 = -0.1$	(l) ARIMA(1, 1, 1)	$\phi_1 = 0.05, \theta_1 = -0.05$

In all cases, the error was white noise with zero mean and unit variance. Time series were grouped into two clusters (stationary and non-stationary) and the clustering evaluation criterion consisted in computing the percentage of successes in the classification. The procedure was replicated $N = 300$ times and the percentage of successes averaged through all the iterations. The results obtained using the complete linkage procedure are shown in Table 1.

For $T = 200$, a group of metrics performed better than the rest (d_{ACFU}, d_{ACFG}, d_{NP} and d_{LNP}), with percentages of success exceeding 80%. The non-parametric measures worked well for low frequency components, with scores around 80%. This is reasonable, since the main differences between the spectra of the two processes type are concentrated in the low frequency band. As expected, d_E and d_P showed the worst performance, with percentages of success below 67% in both cases.

Table 1 Percentage of success in the classification of series (a)–(l) as stationary or non-stationary with $N = 300$ iterations

Measure	T			Measure	T		
	50	200	500		50	200	500
Euclidean distance				*Non-parametric*			
d_E	65.90	66.90	68.20	$d_{W(DLS)}$	73.50	71.00	72.40
Simple, partial autocorrelations				Low freq.	78.06	80.92	82.65
d_{ACFU}	75.40	84.10	83.50	High freq.	62.28	65.92	71.43
d_{ACFG}	76.00	83.20	82.10	$d_{W(LS)}$	67.10	70.70	72.60
d_{PACFU}	74.40	75.00	75.00	Low freq.	71.83	79.81	80.10
d_{PACFG}	74.40	75.00	75.00	High freq.	64.42	67.44	71.94
Periodograms				$d_{W(LK)}$	69.30	71.50	71.90
d_P	66.60	65.80	65.50	Low freq.	75.80	79.50	84.00
d_{LP}	66.00	73.10	74.80	High freq.	63.50	68.10	72.60
d_{NP}	72.00	81.80	82.80	d_{GLK}	63.80	70.72	73.60
d_{LNP}	70.00	84.20	94.40	Low freq.	63.83	79.08	80.8
Low freq.	64.70	73.80	78.60	High freq.	62.50	68.00	72.60
High freq.	69.20	83.80	95.10	d_{CM}	69.50	72.00	74.50
Model-based				Low freq.	75.89	79.89	85.90
d_{PIC}	69.60	74.90	75.00	High freq.	63.11	68.89	72.10
d_M	71.80	75.00	75.00				

T is the length of the series. Low frequencies correspond to ordinates 1 to $\sqrt{T}$. High frequencies to ordinates $\sqrt{T}+1$ to $T/2$

In general, the percentages of success were higher when a larger length ($T = 500$) was used. The performance of the d_{LNP} metric was especially good (95% success). Results with $T = 50$ were rather poor for all measures.

3.2 Clustering of ARMA Time Series

In this case, as in Maharaj (1996), the following models were selected (in all cases the error was again white noise with zero mean and unit variance):

1. AR(1) $\phi_1 = 0.5$
2. MA(1) $\theta_1 = 0.7$
3. AR(2) $\phi_1 = 0.6, \phi_2 = 0.2$
4. MA(2) $\theta_1 = 0.8, \theta_2 = -0.6$
5. ARMA(1, 1) $\phi_1 = 0.8, \theta_1 = 0.2$

Four series of length $T = 200$ were generated from each process and a clustering algorithm was run. One hundred trials were carried out. As a clustering results evaluation criterion the following similarity index was used (see Gavrilov et al. 2000):

$$Sim\,(G, C) = \frac{1}{5}\sum_{i=1}^{5} \max_{1\le j\le k} Sim\left(G_j, C_i\right), \tag{6}$$

Table 2 Clustering of ARMA processes (i)–(v): Cluster similarity evaluation index in (6) for k-cluster solutions

Measure	k		Measure	k	
	4	5		4	5
Euclidean distance			*Model-based*		
d_E	0.457	0.475	d_{PIC}	0.706	0.703
Simple, partial autocorrelations			d_M	0.825	0.807
d_{ACFU}	0.716	0.751	*Non-parametric*		
d_{ACFG}	0.732	0.765	$d_{W(DLS)}$	0.783	0.793
d_{PACFU}	0.820	0.816	$d_{W(LS)}$	0.762	0.774
d_{PACFG}	0.828	0.820	$d_{W(LK)}$	0.778	0.790
Periodograms			d_{GLK}	0.730	0.732
d_P	0.552	0.583	d_{CM}	0.769	0.786
d_{LP}	0.684	0.704			
d_{NP}	0.612	0.648			
d_{LNP}	0.703	0.740			

$T = 200$. $N = 100$. Complete linkage procedure

where $C = \{C_1, \ldots, C_5\}$ and $G = \{G_1, \ldots, G_k\}$, $1 \leq k \leq 5$ are the set of five true clusters and a k-cluster solution, and $Sim\left(G_j, C_i\right) = 2\left|G_j \cap C_i\right|/(\left|G_j\right| + |C_i|)$.

Table 2 provides the average cluster similarity indexes obtained with the complete linkage algorithm. The best scores were obtained with the distances based on the PACF's and the metric d_M, with indexes above 0.8. The d_W-type non-parametric measures are placed in an intermediate position. Among them, the worst performance corresponded to the measure $d_{W(LS)}$. This fact corroborates the asymptotic inefficiency of $d_{W(LS)}$ with respect to both $d_{W(DLS)}$ and $d_{W(LK)}$, as established in Vilar and Pértega (2004). The d_{CM} measure also performed well, while d_{GLK} achieved the worst results among the non-parametric measures.

The rest of the measures performed worse. Note that among these would be included those that best distinguished between stationary and non-stationary processes in our first experiment, namely, d_{ACFU}, d_{ACFG}, d_{NP} and d_{LNP}.

We also analyzed the mean number of times that each of the processes was correctly identified. The measures showing the best behavior, d_M, d_{PACFU}, d_{PACFG} and the non-parametric ones, nearly always correctly identified the MA processes. Series from AR(2) and ARMA(1, 1) models were never well grouped, but this occurred with all the metrics. Series from AR(1) processes were correctly identified 20–40% of the times with the best metrics, except for $d_{W(LS)}$ and d_{GLK}. Concerning the metrics presenting the worst quality indexes, it is interesting to observe their low ability to identify the MA series.

Finally we also considered the mean number of clusters correctly identified at each iteration. The distances based on the PACF's, the Maharaj's distance and the nonparametric distances correctly identified between two and three clusters in each iteration. None of the other metrics were able to correctly identify the mean of two clusters. Furthermore, only the metrics based on the PACF's, Maharaj's measure and the non-parametric measures were able to yield 3 correct clusters at some iteration.

3.3 Clustering of Non-Linear Time Series

In this case, four series of length $T = 200$ were generated from these models:

(1) $X_t = 0.5X_{t-1}I\ (X_{t-1} \leq 0) - 2X_{t-1}I\ (X_{t-1} > 0) + \varepsilon_t$ (TAR model)

(2) $X_t = \left(0.3 - 10\exp\left\{-X_{t-1}^2\right\}\right) X_{t-1} + \varepsilon_t$ (EXPAR model)

(3) $X_t = \varepsilon_t - 0.4\varepsilon_{t-1}$ (Linear MA model)

(4) $X_t = \varepsilon_t - 0.5\varepsilon_{t-1} + 0.8\varepsilon_{t-1}^2$ (Non-linear MA model)

Error processes ε_t were independent zero-mean Gaussian variables with unit variance. These models were used in tests of linearity setting by Tong and Yeung (1991).

The clustering evaluation criterion was again the quality index given in (6). The values of this index (averaged over 100 trials) are shown in Table 3.

As expected, the best results were now attained with the non-parametric dissimilarity measures. All of them led to indexes around 0.9, except for d_{GLK}. The parametric distances d_{PIC} and d_M were affected by the misspecification of the generating models and were placed in an intermediate location together with the autocorrelation-based measures and d_{LP}. The rest of the measures fell far behind. In particular, d_{LNP} yielded a poor quality index of 0.574.

When we analyzed the number of times that each process was correctly identified we observed that the EXPAR and MA processes form the most homogeneous clusters. When the best non-parametric measures ($d_{W(DLS)}$, $d_{W(LK)}$ and d_{CM}) were used, the series from this processes were correctly grouped around 91% of the times. More difficult was to group correctly the series from the TAR and NLMA models. Furthermore, the non-parametric measures led to a complete correct solution nearly 40% of the times. None of the other metrics were able to yield the mean of two correct clusters.

Table 3 Clustering of non-linear processes (1)–(4): cluster similarity evaluation index in (6) for four-cluster solution

Measure	Index	Measure	Index
Euclidean distance		*Model-based*	
d_E	0.537	d_{PIC}	0.769
Simple, partial autocorrelations		d_M	0.781
d_{ACFU}	0.752	*Non-parametric*	
d_{ACFG}	0.777	$d_{W(DLS)}$	0.920
d_{PACFU}	0.784	$d_{W(LS)}$	0.895
d_{PACFG}	0.795	$d_{W(LK)}$	0.912
Periodograms		d_{GLK}	0.818
d_P	0.485	d_{CM}	0.913
d_{LP}	0.786		
d_{NP}	0.576		
d_{LNP}	0.574		

$T = 200$. $N = 100$. Complete linkage procedure

4 Concluding Remarks

Simulation results shown that the performance of a dissimilarity metric depends on the type of processes subjected to clustering. For example, both the metric based on the log-normalized periodograms and the metric based on the simple autocorrelation functions yielded the highest success rates for discriminating between stationary and non-stationary processes. These two measures turned out to be among the worst ones when they were used to cluster ARMA and non-linear processes. In a similar way, model-based metrics can suffer from the effects of a misspecification.

Among the measures examined, we included up to five measures based on non-parametric criteria. Such as we expected, these non-parametric measures performed reasonably well in the three clustering setups considered and, in fact, they were the only ones showing this robustness. Specifically, all of them provided substantially better results than the rest in the clustering of non-linear processes, they presented results very close to the best ones in clustering of ARMA processes, and their success rates in distinguishing between stationary and non-stationary processes were fairly competitive when these measures were evaluated in the low frequency range. From this group of non-parametric measures, $d_{W(DLS)}$, $d_{W(LK)}$ and d_{CM} showed the best behavior, and therefore, they could be declared as the "winners" in our simulation study.

Acknowledgements This work was supported by the Ministerio de Ciencia e Innovación, Grant MTM2008-00166 (ERDF included).

References

Caiado, J., Crato, N., & Peña, D. (2006). A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis, 50*, 2668–2684.

Fan, J., & Kreutzberger, E. (1998). Automatic local smoothing for spectral density estimation. *Scandinavian Journal of Statistics, 25*, 359–369.

Fan, J., & Zhang, W. (2004). Generalised likelihood ratio tests for spectral density. *Biometrika, 91*, 195–209.

Galeano, P., & Peña, D. (2000). Multivariate analysis in vector time series. *Resenhas, 4*, 383–403.

Gavrilov, M., Anguelov, D., Indyk, P., & Motwani, R. (2000). Mining the stock market (extended abstract): which measure is best? *Proceedings of the 6th Internatinal Conference on Knowledge Discovery and Data Mining (KDD)*. August 20–23, pp. 487–496, Boston, MA, USA.

Kakizawa, Y., Shumway, R. H., & Taniguchi, M. (1998). Discrimination and clustering for multivariate time series. *Journal of the American Statistical Association, 93*, 328–340.

Liao, T. W. (2005). Clustering of time series data: a survey. *Pattern Recognition, 38*, 1857–1874.

Maharaj, E. A. (1996). A significance test for classifying ARMA models. *Journal of Statistical Computation and Simulation, 54*, 305–331.

Maharaj, E. A. (2000). Clusters of time series. *Journal of Classification, 17*, 297–314.

Piccolo, D. (1990). A distance measure for classifying arima models. *Journal of Time Series Analysis, 11*, 153–164.

Tong, H., & Yeung, I. (1991). On tests for self-exciting threshold autoregressive type non-linearity in partially observed time series. *Applied Statistics, 40*, 43–62.

Vilar, J. A., & Pértega, S. (2004). Discriminant and cluster analysis for Gaussian stationary processes: Local linear fitting approach. *Journal of Nonparametric Statistics 16*, 443–462.
Vilar, J. M., Vilar, J. A., & Pértega, S. (2009) Classifying time series data: A nonparametric approach. *Journal of Classification*, 2009 (to appear).

Finite Dimensional Representation of Functional Data with Applications

Alberto Muñoz and Javier González

Abstract Most algorithms in statistics are designed to work with vectors of small or moderate dimension, and the performance of these algorithms decreases when dealing with very high dimensional data as functional data are. In this work we propose a functional analysis technique to obtain appropriate finite-dimensional representations of functional data for pattern recognition purposes. To this aim, we project the available functional data samples onto finite dimensional function spaces generated by the eigenfunctions of suitable Mercer kernels. We demonstrate some theoretical properties of the proposed method and the advantages of the proposed representations in several tasks using simulated and real functional data sets.

1 Introduction

Functional data sets are characterized by their very high (or intrinsically infinite) dimensionality (Ramsay and Silverman 2006). Functional data examples arise naturally in fields such as chemometrics, climatology, etc. Functional Data Analysis (FDA) methods study the generalization of multivariate techniques for the case in which the data are functions. A usual technique to apply statistical procedures to FDA data sets is to project functional data points onto some given finite-dimensional function subspace (Ramsay and Silverman 2006). In this paper we choose to work with Reproducing Kernel Hilbert Spaces (RKHS) as reference functional space. In particular, we propose a finite-dimensional representation for functional data based on a particular projection of the original functions onto the subspace generated by the eigenfunctions of a given RKHS kernel. In Sect. 2 we formulate the functional data representation in the context of regularization theory for the Square Loss function. In Sect. 3 we exemplify the operation of the technique in some real and simulated examples. Section 4 contains the conclusions.

J. González (✉)
Universidad Carlos III de Madrid, c/ Madrid 126, 28903 Getafe, Spain
e-mail: javier.gonzalez@uc3m.es

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_16,

2 Representing Functional Data in a Reproducing Kernel Hilbert Space

The goal is to transform each functional datum (usually a curve) into a point of a given RKHS. Let $\{\hat{c}_1, \ldots, \hat{c}_m\}$ denote the available sample of curves. Each sample curve $\hat{c}_l$ is given by a data set $\{(\mathbf{x}_i, \mathbf{y}_{il}) \in X \times Y\}_{i=1}^n$ where X represents the input space and, usually $Y = \mathbb{R}^n$. Here $n = 1$ (we deal with curves). We assume that, for each $\hat{c}_l$, there exists a continuous function $c_l : X \longrightarrow Y$ such that $E[y_l|\mathbf{x}] = c_l(\mathbf{x})$ (with respect to some probability measure). Thus $\hat{c}_l$ is the sample version of c_l. We assume that the $\mathbf{x}_i$ are common for all the curves, as it is the habitual case in the literature (Ramsay and Silverman 2006).

A Hilbert function space H is a RKHS when there exists a symmetric positive definite function $K : X \times X \to \mathbb{R}$ such that the elements of H can be expressed as finite linear combinations of the form $h = \sum_s \lambda_s K(x_s, \cdot)$ where $\lambda_s \in \mathbb{R}$ and $x_s \in X$. K is called Mercer Kernel or reproducing kernel for H (Aroszajn 1950) and H is denoted by H_K. For more details on RKHSs see Aroszajn (1950); Cucker and Smale (2002); Wahba (1990); Moguerza and Muõz (2006).

Consider the linear integral operator T_K associated to K defined by $T_K(f) = \int_X K(\cdot, s) f(s) ds$. If we impose that $\iint K^2(x, y) dx dy < \infty$, then T_K has a countable sequence of eigenvalues $\{\lambda_j\}$ and (orthonormal) eigenfunctions $\{\phi_j\}$ and K can be expressed as $K(x, y) = \sum_j \lambda_j \phi_j(x) \phi_j(y)$ (where the convergence is absolute and uniform).

Given a function f in a general function space, it will be projected onto H_K by means of the operator T_K: the projection f^* will belong to the range of T_K: $f^* = T_K(f)$. Applying the Spectral Theorem to T_K we get:

$$f^* = T_K(f) = \sum_j \lambda_j \langle f, \phi_j \rangle \phi_j \tag{1}$$

Next we want to obtain c_l^* for each c_l (the function corresponding to the sample functional data point $\hat{c}_l \equiv \{(\mathbf{x}_i, y_{il}) \in X \times Y\}_{i=1}^n$). To find the coefficients of c_l^* in (1), we apply regularization theory (Chen and Haykin 2002): c_l^* will be the function that solves the following variational problem (Cucker and Smale 2002; Moguerza and Muõz 2006):

$$\arg \min_{c \in H_K} \frac{1}{n} \sum_{i=1}^{n} L(y_{il}, c(\mathbf{x}_i)) + \gamma \|c\|_K^2 . \tag{2}$$

where $\gamma > 0$, $\|c\|_K$ is the norm of the function c in H_K, and the loss function is $L(y_{il}, c(\mathbf{x}_i)) = (c(\mathbf{x}_i)) - y_{il})^2$. The functional in (2) measures the trade-off between the fitness of the function to the data and the complexity of the solution (measured by $\|c\|_K^2$) and problem (2) can be solved applying the following theorem, known as the Representer Theorem. For details, proofs and generalizations, refer to Kimeldorf and Wahba (1971); Schölkopf and Herbrich (2001); Cox and O'Sullivan (1990).

Theorem 1 (Representer Theorem). *Given the sample curve $\hat{c}_l$, defined by the set $\{(x_i, y_{il}) \in X \times Y\}_{i=1}^n$, the solution c_l^* to the functional optimization problem (2) exists, is unique and admits a representation of the form*

$$c_l^*(x) = \sum_{i=1}^{n} \alpha_{il} K(x_i, x) \quad \forall x \in X, \text{ where } \alpha_i \in \mathbb{R}. \tag{3}$$

In order to obtain the coefficients α_i we have to solve the linear system $(\gamma n I_n + K_S)\alpha_l = y_l$, where $K_S = (K(x_i, x_j))_{i,j}$.

By solving this linear we get a closed expression for c_l^*, the minimizer of problem (2). Next we define two functional data representations starting from (3).

2.1 Functional Data Projections onto the Eigenfunctions Space

The minimization of the risk functional (2) gives the projected points $c_1^*, \ldots c_m^*$ in H_K corresponding to the original curves $\{\hat{c}_1, \ldots, \hat{c}_m\}$. Equation (3) provides a concrete finite-dimensional representation for each curve $\hat{c}_l$, namely the set of coefficients $\alpha_{1l}, \ldots, \alpha_{nl}$. Unluckily, as we will see right away, this representation has a serious inconvenience: if the sample is slightly different, say $(\mathbf{x}_i')$, then it may happen that the corresponding y_{il}' are quite different, making the representation system not valid for pattern recognition purposes:

Theorem 2. *Let c be a curve, whose sample version is $\{(\boldsymbol{x}_i, y_i) \in X \times Y\}_{i=1}^n$. The representation of c in terms of the vector $(\alpha_1, \ldots, \alpha_n)$ is not continuous, where $\{\alpha_i\}_{i=1}^n$ are the coefficients of c^* in (3): $c^*(\boldsymbol{x}) = \sum_{i=1}^n \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x})$.*

Proof. The number of non null terms in the sum $K(\mathbf{x}, \mathbf{y}) = \sum_i \lambda_i \phi^T(\mathbf{x})\phi(\mathbf{y})$ is $d = \min(n, r(T_K))$, where $r(T_K)$ is the rank of the operator T_K ($\lambda_j = 0$ for $j > r(T_K)$).

$$\begin{aligned} c^*(\mathbf{x}) &= \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x}) = \sum_{i=1}^{n} \alpha_i \left(\sum_{j=1}^{d} \lambda_j \phi_j(\mathbf{x}_i) \phi_j(\mathbf{x}) \right) \\ &= \sum_{j=1}^{d} \lambda_j \left(\sum_{i=1}^{n} \alpha_i \phi_j(\mathbf{x}_i) \right) \phi_j(\mathbf{x}) \end{aligned} \tag{4}$$

On the other hand, by (1) $c^*(\mathbf{x}) = \sum_{j=1}^d \lambda_j \langle c, \phi_j \rangle \phi_j(\mathbf{x})$. Equating to (4) and being the $\{\phi_j\}$ a basis for H_K we get: $\langle g, \phi_j \rangle = \sum_i \alpha_i \phi_j(\mathbf{x}_i) = \langle \alpha, \phi_j \rangle$. Therefore, for any set $\alpha' = (\alpha_1', \ldots, \alpha_n')$ such that $\langle \alpha', \phi_j \rangle = \langle \alpha, \phi_j \rangle = \langle g, \phi_j \rangle$ we will have that $\sum_{i=1}^n \alpha_i' k(\mathbf{x}_i, \mathbf{x}) = c^*(\mathbf{x})$. Now, given the sample curve $c \equiv \{(\mathbf{x}_i, y_i) \in X \times Y\}_{i=1}^n$, consider a 'close' curve $c^\epsilon \equiv \{(\mathbf{x}_i^\epsilon, y_i^\epsilon) \in X \times Y\}_{i=1}^n$, such that $d(\mathbf{x}, \mathbf{x}^\epsilon) < \epsilon$. Denote

by (α^ϵ) the representation corresponding to c^ϵ. Given that $c^\epsilon(\mathbf{x}) \simeq c^*(\mathbf{x})$ (because of the continuity of c), and using (4) it will happen that $\langle \alpha^\epsilon, \phi_j \rangle \simeq \langle \alpha, \phi_j \rangle$ and, nevertheless, by the previous reasoning, α^ϵ and α can be quite different. In practice, this situation can arise when the matrix $(\gamma n I_n + K_S)$ is close to be singular: a small change in the sample can cause a large change in the solution of the corresponding linear system.

The next theorem specifies our concrete proposal to obtain finite-dimensional representations for functional data.

Theorem 3. *Let c be a curve, whose sample version is $\hat{c} \equiv \{(\boldsymbol{x}_i, y_i) \in X \times Y\}_{i=1}^n$. Consider the functional representation for c given by $(\lambda_1^*, \ldots, \lambda_d^*)$, where*

$$\hat{\lambda}_j^* = \sum_{i=1}^{n} \hat{\lambda}_j \alpha_i \hat{\phi}_{ji} \ , \tag{5}$$

α_i are given by (3), $\hat{\lambda}_j$ is the eigenvalue corresponding to the eigenvector $\hat{\phi}_j$ of the matrix $K_S = \left(K(\boldsymbol{x}_i, \boldsymbol{x}_j)\right)_{i,j}$, and $d = \min(n, r(K_S))$. This functional representation is continuous with respect to the input variables.

Proof. In the ideal case where we know the expression for both the eigenfunctions and eigenvalues of the kernel function K, $\lambda_j^* = \sum_{i=1}^n \lambda_j \alpha_i \phi_j(\mathbf{x}_i)$. However, often we only know the matrix K_S, obtained by evaluating the kernel at the sample, and we can not know the real eigenvalues λ_j and their corresponding eigenfunctions ϕ_j. We will prove the theorem for the representation given by $\sum_{i=1}^n \lambda_j \alpha_i \phi_j(\mathbf{x}_i)$, and then we show that $\sum_{i=1}^n \hat{\lambda}_j \alpha_i \hat{\phi}_{ji}$ converges to that value. First we show that $\sum_{j=1}^d \lambda_j^* \phi_j(\mathbf{x})$ gives the value of $c^*(\mathbf{x})$:

$$\begin{aligned} \sum_{j=1}^{d} \lambda_j^* \phi_j(\mathbf{x}) &= \sum_{j=1}^{d} \left(\lambda_j \sum_{i=1}^{n} \alpha_i \phi_j(\mathbf{x}_i) \right) \phi_j(\mathbf{x}) = \sum_{j=1}^{d} \lambda_j \left(\sum_{i=1}^{n} \alpha_i \phi_j(\mathbf{x}_i) \right) \phi_j(\mathbf{x}) \\ &= \sum_{i=1}^{n} \alpha_i \left(\sum_{j=1}^{d} \lambda_j \phi_j(\mathbf{x}_i) \phi_j(\mathbf{x}) \right) \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x}) = c^*(\mathbf{x}) \end{aligned} \tag{6}$$

Given the sample curve c and a 'close' curve c^ϵ, and using the same notation as in Theorem 2, if $d(x, x^\epsilon) < \epsilon$ then $c(\mathbf{x}) \simeq c^\epsilon(\mathbf{x})$ and given that the ϕ_j are a basis for H_K, it must happen that $\lambda_j^* \simeq \lambda_j^{*\epsilon}$. To end the proof, we only need to show that the eigenvalues and eigenvectors of K_S converge, respectively, to the eigenvalues and eigenfunctions of T_K: $\hat{\lambda}_j \longrightarrow \lambda_j$ and $\hat{\phi} \longrightarrow \phi$. And this is the case because this convergence holds always for positive-definite matrices, including kernel functions (see Schlesinger 1957). For more specific theorems restricted to the context of kernel functions, see Bengio et al. (2004).

3 Experiments

As a first example we consider two similar functional data curves to illustrate the behavior of the Kernel expansion and the RKHS representation system. The two curves are temperatures curves corresponding to daily series averaged over the period from 1960 to 1994 in Canada (Ramsay and Silverman 2006, Chap. 1), and correspond to the cities "St. Johns" and "Halifax". We consider the kernel $K(\mathbf{x}, \mathbf{y}) = e^{-\sigma \|\mathbf{x}-\mathbf{y}\|^2}$ with $\sigma = 0.4$ and $\gamma = 1$ and obtain the kernel expansion [given in (3)] and the RKHS representation [given in (5)] for both curves. Figure 1, left (upper and lower), shows the curves and their projections onto the function space H_K generated by the eigenfunctions of K. The two central plots in Fig. 1 show the kernel expansion representation for both curves and it is apparent they are quite different, despite the fact the two curves are similar. Figure 1, right, shows the RKHS representations for both curves and now they look similar, in agreement with Theorem 3. In addition, we can see that the (λ_j^*) are representing the curves in a four-dimensional space, which agrees with the result obtained by the dimensionality test proposed in Hall and Vial (2006). We can therefore conclude that the RKHS representation is robust against the presence of noise in the data.

3.1 RKHS Projections Versus PCA Projections

In Statistics it is usual to reduce the dimension of high dimensional data before affording cluster or classification tasks. In FDA this is achieved by using the Functional Principal Components (FPCA) (Ramsay and Silverman 2006). As in the

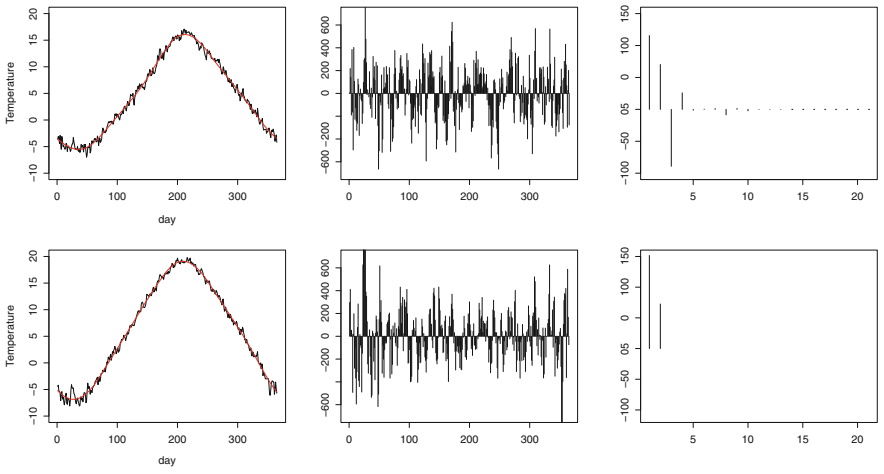


Fig. 1 Two Canadian curves, and their kernel expansion and RKHS representations

multivariate case, this technique make use of the data covariance function to determine the subspace onto the data are projected. This subspace is spanned by the data covariance eigenfunctions and is always a RKHS (see Rakotomamonjy and Canu 2005). Within this setting, FPCA can be considered a particular case of our methodology.

The election of the data covariance matrix as kernel K in Theorem 3 is justified in certain theoretical cases (see James and Sugar 2003). In practice, more general kernels can be considered. The next example illustrates this in a clustering problem. Consider two families of 10 dimensional curves sampled at 500 points:

- Class 1: $c(x) = \sum_{j=1}^{10} a_j \phi_j(x) = \sin(j\pi x)$, where $a_i \sim N_{10}(\mu_1, \Sigma)$
- Class 2: $c(x) = \sum_{j=1}^{10} b_j \phi_j(x) = \sin(j\pi x)$, where $b_j \sim N_{10}(\mu_2, \Sigma)$

with $x \in [0, 1]$ and for $\mu_1 = (8, 8, 1, 2, 3, 4, 5, 6, 7, 8)$, $\mu_2 = (-8, -8, 1, 2, 3, 4, 5, 6, 7, 8)$, and $\Sigma = diag(1, 150, 150, 10, 10, 10, 10, 10, 10, 10)$. For our experiment, we generated 50 curves of each family (see Fig. 2).

We are going to compare the RKHS representation system with the data covariance and with a generalized covariance: an exponential kernel. First we try to separate (automatically) the curves using row data. We performed 10 runs of a k-means algorithm (with two centroids) and a hierarchical cluster by using the Ward method. The misclassification errors were 25.2% and 24% respectively. By using FPCA, the first two principal components explain over 80% of the variability. This two components are plotted in Fig. 3 (left). Applying the two previous cluster procedures over this projection we obtain misclassification errors of 15% (for the k-means) and 18% (for the hierarchical cluster). The dimension reduction improves the results but a large number of curves is still assigned to wrong families. On the other hand, if the two first projections are achieved by using the kernel $K(\mathbf{x}, \mathbf{y}) = e^{-10\|\mathbf{x}-\mathbf{y}\|^2}$ with regularization parameter $\gamma = 1$, (see Fig. 3) 0% of errors are obtained with both cluster algorithms.

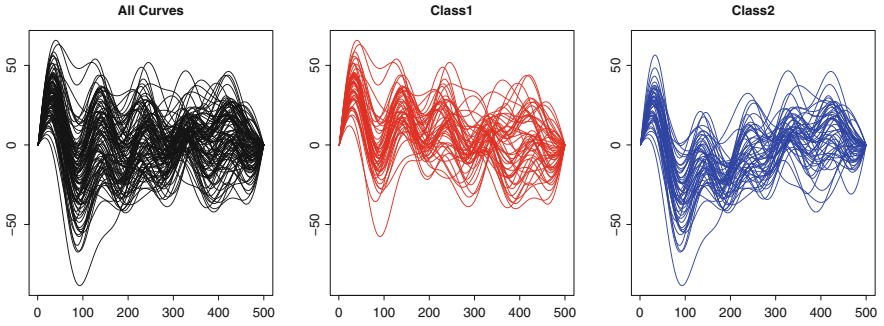


Fig. 2 *Left*: all curves together. *Center*: Class 1 curves. *Right*: Class 2 curves

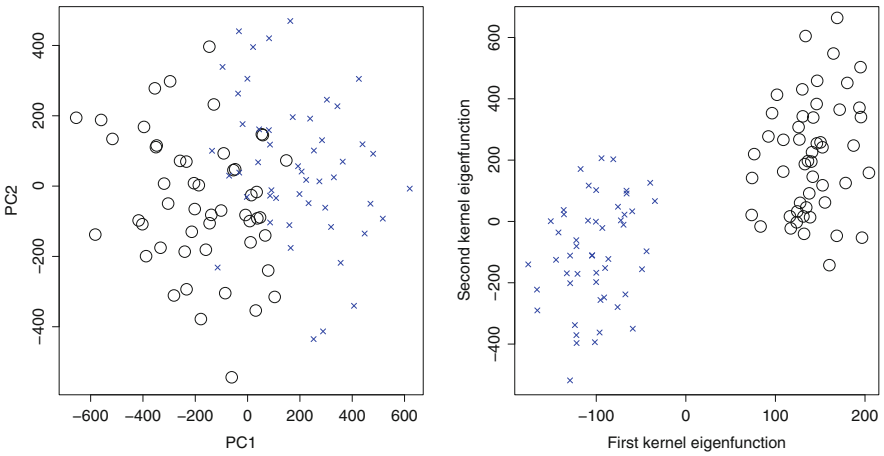


Fig. 3 Two first FPCA projection (*left*) and RKHS projections (*right*)

3.2 Classification Example

The data set is made up of 215 observations in the near infrared absorbance spectrum of a meat sample. The classes are determined by the fat content: class 1, more that 20%, class 2, less than 20%. Following Rossi and Villa (2006), we used the second derivative of the data. In Fig. 4 both classes are shown for the original curves and for the second derivatives. In order to test the RKHS projections in this classification example, we divide the sample in training data (80% of the observations) and testing (20%). We project each data set onto two different RKHSs: we considered in the covariance matrix as kernel in the regularization process (K_1) and $K_2(\mathbf{x}, \mathbf{y}) = e^{-0.1\|\mathbf{x}-\mathbf{y}\|^2}$. The regularization parameter was set, in all cases, $\gamma = 1$. The steps for our curve classification proposal are: (1) Project the data onto a RKHS. That is, estimate the representation (λ_j^*). (2) Classify the data with a linear Support Vector Machine (SVM) with $C = 1$ by setting $\widehat{c_l} \equiv \{\lambda_{lj}^*\}$.

The number of components of the projection used to classify is determined by cross validation. The election of the SVM is twofold. It is statistically consistent and it is proven to have a good performance in real examples (Moguerza and Muõz 2006). This election is, of course, generalizable to any other classification technique. We compared the RKHS projections misclassification errors with two specific techniques designed to deal with functional data: the P-spline signal regression, developed by Marx and Eilers (1999) the and MPLSR/Knn developed by Ferraty and Vieu (2003). Second derivative metric of the Tecator was selected following Ferraty and Vieu (2003)

Results are shown in Table 1. It is clear that the classification rates achieved by the RKHS projections are the most competitive. The RBF projection achieves an misclassification error of 1.4%. However, in this case this error rate is outperformed by the covariance projection (0.9%) showing the utility of this election in some examples.

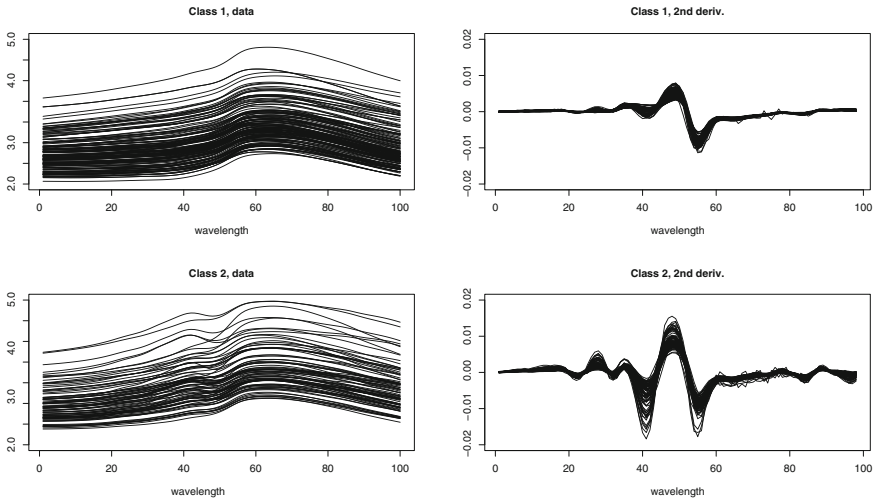


Fig. 4 Two classes of curves of the Tecator data set. Original curves and second derivatives are shown

Table 1 Comparative results after 100 runs

Method	RKHS K_1	RKHS K_2	*PSR*	$MPLSR_{(d^2)}$
dim	4	3	–	–
Test Error	0.0097	0.0145	0.0731	0.0218
Std. Dev	0.0014	0.0017	0.0031	0.0020

The values of *dim* in the RKHS projections was selected by cross validation

4 Conclusions

In this work we have proposed a system to represent functional data, by projecting the original functions onto the eigenfunctions of a Mercer kernel with the aid of regularization theory. A main advantage is that we do not have to specify the basis of eigenfunctions, but we can concentrate in the kernel, following the general philosophy of kernel methods. The proposed representation works well in the experiments. We have checked, in a real temperature example, how it is able to capture the interesting features of functional data set of curves. In addition, two real examples were analyzed following our methodology obtaining the best error rates.

References

Aroszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society, 68*(3), 337–404.

Bengio, Y., Delalleau, O., Le Roux, N., Paiement, J.-F., Vincent, P., & Ouimet, M. (2004). Learning eigenfunctions links spectral embedding and kernel PCA. *Neural Computation, 16*, 2197–2219.

Chen, Z., & Haykin, S. (2002). On different facets of regularization theory. *Neural Computation, 14*, 2791–2846.

Cox, D., & O'Sullivan, F. (2004). Asymptotic analysis and penalized likelihood and related estimators. *Annals of Statistics, 18*, 1676–1695.

Cucker, F., & Smale, S. (2002). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society, 39*(1), 1–49.

Ferraty, F., & Vieu, P. (2003). Curves discrimination: A nonparametric functional approach. *Computational Statistics & Data Analysis, 44*, 161–173.

Hall, P., & Vial, C. (2006). Assessing the finite dimensionality of functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68*(4), 689–705.

James, G. M., & Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association, 98*.

Kimeldorf, G. S., & Wahba, G. (1971). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics, 2*, 495–502.

Marx, B., & Eilers, P. (1999). Generalized linear regression on sampled signals and curves: A P-spline approach. *Technometrics 41*, 113.

Moguerza, J. M., & Muõz, A. (2006). Support vector machines with applications. *Statistical Science, 21*(3), 322–357.

Rakotomamonjy, A., & Canu, S. (2005). Frames, reproducing kernels, regularization and learning. *Journal of Machine Learning Research 6*, 1485–1515.

Ramsay, J. O., & Silverman, B. W. (2006). Functional data analysis (2nd ed.). New York: Springer.

Rossi, F., & Villa, N. (2006). Support vector machine for functional data classification. Neurocomputing 69, 730–742.

Schlesinger, S. (1957). Approximating eigenvalues and eigenfunctions of symmetric kernels. *Journal of the Society for Industrial and Applied Mathematics, 6*(1), 1–14.

Schölkopf, B., Herbrich, R., Smola, A. J., & Williamson, R. C. (2001). A generalized representer theorem. *Lecture Notes in Artificial Intelligence, 2111*, 416–426, Springer.

Wahba, G. (1990). Spline models for observational data. *Series in Applied Mathematics, 59*, Philadelphia: SIAM.

Clustering Spatio-Functional Data: A Model Based Approach

Elvira Romano, Antonio Balzanella, and Rosanna Verde

Abstract In many environmental sciences, such as, in agronomy, in metereology, in oceanography, data analysis has to take into account both spatial and functional components. In this paper we present a strategy for clustering spatio-functional data. The proposed methodology is based on concepts of spatial statistics theory, such as variogram and covariogram when data are curves. Moreover a summarizing spatio-functional model for each cluster is obtained.

The assessment of the method is carried out with a study on real data.

1 Introduction and Problematic

Functional data analysis is about the analysis of information on curves or functions (Ramsay and Silverman 2005). There is a large number of applicative fields, such as agronomy and climatology, where functional data are observed in spatial varying way. This has led, in the last year, to the development of a relatively new branch of statistics: Spatial Functional Data Analysis (Ramsay 2008). The literature in this framework is not extensive, the problem more considered, that is the also more frequent in real studies, is the analysis of functional data presenting spatial dependence.

Three classic types of spatial data structures usually referred as geostatistical data, point patterns and areal data can be combined with functional data (Delicado et al. 2009).

In this work we focus on geostatistical data where a finite sample of measurements relating to an underlying spatially continuous phenomenon is observed. In such context, we introduce a clustering strategy where measurements are functional

E. Romano (✉)
Facoltá di Studi Politici, Dipartimento di Studi Europei e Mediterranei, Seconda Universitá degli Studi di Napoli, Via del Setificio 15, 81100 Caserta
e-mail: elvira.romano@unina2.it

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_17,

data. We take into account both the spatial and functional information, using tools from Geostatistics and Functional Data analysis (Ramsay and Silverman 2005).

Two usual ways for clustering such kind of data exist. The first one, follows the Functional Data Analysis approach and consists in considering only the functional nature of data (e.g. Abraham et al. 2005; Heckman and Zamar 2000; James and Sugar 2005; Romano 2006), the second way, takes into consideration the spatial location to perform the analysis of data (Blekas et al. 2007).

The proposed algorithm is a special case of Dynamic Clustering Algorithm (e.g., Diday 1971) which finds an optimal partition of the functional data and a set of representation functions. It is based on an optimization problem that minimizes the spatial variability among the curves in each cluster. The representation functions are a spatio-functional linear models for delimited spatial area selected according to the clustering strategy.

The paper is organized as follows: in Sect. 2 we present the data structure, in Sect. 3 we recall the Dynamic Clustering algorithm, in Sect. 4 we introduce our strategy, in Sect. 5 we present an application on real data. Finally conclusions and discussions of future work are given in Sect. 5.

2 The Spatio-Functional Data

Let $\{\boldsymbol{\chi}_s(t) : t \in T, s \in D \subset R^d\}$ be a random field where the set $D \subset R^d$ is a fixed subset of R^d with positive volume. $\boldsymbol{\chi}_s$ is a functional variable defined on some compact set T of R for any $s \in D$. We assume to observe a sample of curves $\chi_{s_1}(t), \ldots, \chi_{s_i}(t), \ldots, \chi_{s_n}(t)$ for $t \in T$ where s_i is a generic data location in the d-dimensional Euclidean space. We assume for each t that we have a second order stationary and isotropic random process, that is, the mean and variance functions are constant and the covariance depends only on the distance between sampling sites.

Formally, we assume that:

- $E(\boldsymbol{\chi}_s(t)) = m(t)$, for all $t \in T,\ s \in D$.
- $V(\boldsymbol{\chi}_s(t)) = \sigma^2$, for all $t \in T,\ s \in D$.
- $Cov(\chi_{s_i}(t), \chi_{s_j}(t)) = C(h,t)$ where $h_{ij} = \|s_i - s_j\|$ and all $s_i, s_j \in D$
- $\frac{1}{2}V(\chi_{s_i}(t), \chi_{s_j}(t)) = \gamma(h,t) = \gamma_{s_i s_j}(t)$ where $h_{ij} = \|s_i - s_j\|$ and all $s_i, s_j \in D$.

The function $\gamma(h,t)$ as function of h is called semivariogram of $\chi(t)$.

Our proposal is to partition the random field $\{\boldsymbol{\chi}_s : s \in D \subset R^d\}$ into a set of C clusters such that the obtained clusters contain spatially related curves. In the following, before to introduce the proposed strategy, we shortly recall the general scheme of the Clustering Algorithm.

3 Dynamic Clustering Algorithm

Dynamic clustering algorithm (DCA) or *Nueés Dynamiques*is an unsupervised batch training algorithm. Like in the classical clustering techniques the aim is to find groups that are internally dense and sparsely connected with the others. Let E be a set of n objects, it looks for the partition $P = \{P_1, \ldots, P_C\} \in \mathcal{P}_c$ (where $\mathcal{P}_c$ is the family of all the partition $\{P_1, \ldots, P_c\} \in \mathcal{P}_C$ in C clusters) and a set $G = \{g_1, \ldots, g_C\} \in \mathcal{G}_c$ (where $\mathcal{G}_c$ is the family of all admissible representation of C clusters prototypes) such that the criterion of best fitting between G and P is minimized

$$\Delta(P^*, G^*) = Min\{\Delta(P, G) \mid P \in \mathcal{P}_c, \in \mathcal{G}_c\} \tag{1}$$

This criterion is usually an additive function on the C clusters and on the n elements of E, it is defined as:

$$\Delta(P, G) = \sum_{c=1}^{C} \sum_{i \in P_c} \delta(P_c, g_c) \tag{2}$$

where $\delta(P_c, g_c)$ is a function which measures how well the prototype g_c represents the characteristics of objects of the cluster and it can be usually interpreted as an heterogeneity or a dissimilarity measure of goodness of fit between g_c and P_c. Starting from an initial set of clusters, the method applies a *representation* function and an *allocation* function, and these steps are iterated until the convergence.

4 Dynamic Clustering for Spatio-Functional Data

Since we deal with spatio-functional data, the criterion to optimize becomes:

$$\Delta(P, G) = \sum_{c=1}^{C} \sum_{i \in P_c} \int_T V\left(\chi_{s_i}(t) - \chi_{s_c}(t)\right) dt \tag{3}$$

where $V\left(\chi_{s_i}(t) - \chi_{s_c}(t)\right)$ is the spatial variability.

Since we assumed that data are generated from a functional linear concurrent model (Ramsay and Silverman 2005) the criterion can be written

$$\Delta(P, G) = \sum_{c=1}^{C} \sum_{i \in P_c} \int_T V\left(\chi_{s_i}(t) - \sum_{i=1}^{n_c} \lambda_i \chi_{s_i}(t)\right) dt \qquad u.c. \sum_{i=1}^{n_c} \lambda_i = 1 \tag{4}$$

where the prototype $\chi_{s_c} = \sum_{i=1}^{n_c} \lambda_i \chi_{s_i}(t)$ is an ordinary kriging predictor for curves in the cluster c at the optimal spatial location s_c and the kriging coefficients λ_i represent the contribute of each curve to the prototype estimation.

According to the proposed criterion, the parameters to estimate are: the kriging coefficients, the spatial location of the prototypes, the spatial variance for each cluster.

For a fixed value of s_c, the estimation of the n_c kriging coefficients λ_i of each cluster is a constrained minimization problem, due to the unbiasedness constraint. So it is necessary to solve a linear system by means of Langrange multiplier method. In this paper we refer to the method proposed in Delicado et al. (2007), that in matrix notation, can be seen as the minimization of trace of the mean-squared prediction error matrix in the functional setting.

According to this approach a global uncertainty measure associated to the trace-semivariogram $\int_T \gamma_{s_i,s_j(t)}dt$, is given by:

$$\int_T V\left(\chi_{s_i}(t) - \sum_{i=1}^{n_c}\lambda_i\chi_{s_i}(t)\right)dt = \sum_{i=1}^{n_c}\lambda_i\int_T \gamma_{s_i,s_c(t)}dt - \mu \quad u.c. \sum_{i=1}^{n_c}\lambda_i = 1 \tag{5}$$

It is an integrated version of the classical pointwise prediction variance of ordinary kriging and gives indication on the goodness of fit of the predicted model.

In the ordinary kriging for functional data the problem is to obtain an estimate of a curve in an unsampled location. In our case we aim to obtain not only the prediction of the curve, but also the best representative location. We suppose that the prototype of each cluster is located on a cell of a regular spaced grid built starting from locations the functional observations. In this sense the location is a parameter to estimate and the objective function may have several local minima correspondent to different local kriging. We propose to solve this problem evaluating for each cluster, the local kriging on the cells of the spatial grid. The prototype χ_{s_c} and its location s_c is obtained as the best predictor in terms of spatio-functional fitting (5) among the set of estimates on the unsumpled locations of the grid.

Once we have estimated the prototypes we allocate each new curve to the cluster according to the following *allocation* function:

$$\kappa = \chi_{\int} \longmapsto \mathcal{P}_c \tag{6}$$

It allows to assign χ_{s_i} to cluster c of P_c $\kappa(G) = P = \{P_1, \ldots, P_C\}$, according to the minimum-spatial variability rule:

$$P_c := \{i \in \boldsymbol{\chi}_s : \delta(\{i\}, \chi_{s_c}) \leq \delta(\{i\}, \chi_{s_{c*}}) \; for \; 1 \leq c^* \leq C\} \tag{7}$$

with:

$$\delta(\{i\}, \chi_{s_c}) = \frac{1}{\lambda_\alpha} d^2\left(\chi_{s_i}(t); \chi_{s_c}(t)\right) \tag{8}$$

where $d^2\left(\chi_{s_i}(t); \chi_{s_c}(t)\right)$ is the Euclidean distance and λ_α is the kriging coefficient or weight of the curve $\chi_{s_\alpha}(t)$ such that $|s_\alpha - s_c| \cong |s_i - s_c|$.

Applying iteratively the representation function followed by the allocation function the algorithm converges to a stationary value. The convergence of the criterion

is guaranteed by the consistency between the way to represent the classes and the proprieties of the allocation function.

5 Analysis of a Real Dataset: Sea Temperature of the Italian Coast

To evaluate the effectiveness of the proposed strategy, a test have been performed on a real dataset which stores the sea temperature along several locations of the Italian Coast (see http://www.mareografico.it).

The mareographic Network is composed by 26 survey stations undistributed across the italian territory and located mainly within the harbours of Trieste, Venezia Lido, Ancona, Ravenna, Pescara, Ortona, Isole Tremiti, Vieste, Bari, Otranto, Taranto, Crotone, Reggio Calabria, Messina, Catania, Porto Empedocle, Lampedusa, Palermo, Palinuro, Salerno, Napoli, Cagliari, Carloforte, Porto Torres, Civitavecchia, Livorno, Genova ed Imperia. Several kind of meteorological phenomena are recorded over the time, we use temperature curves of water.

For each location, we have a curve recorded in a period of 2 weeks with spatial coordinates (s_x, s_y) correspondent respectively to the latitude and longitude. A sample of the data can be observed in Fig. 1.

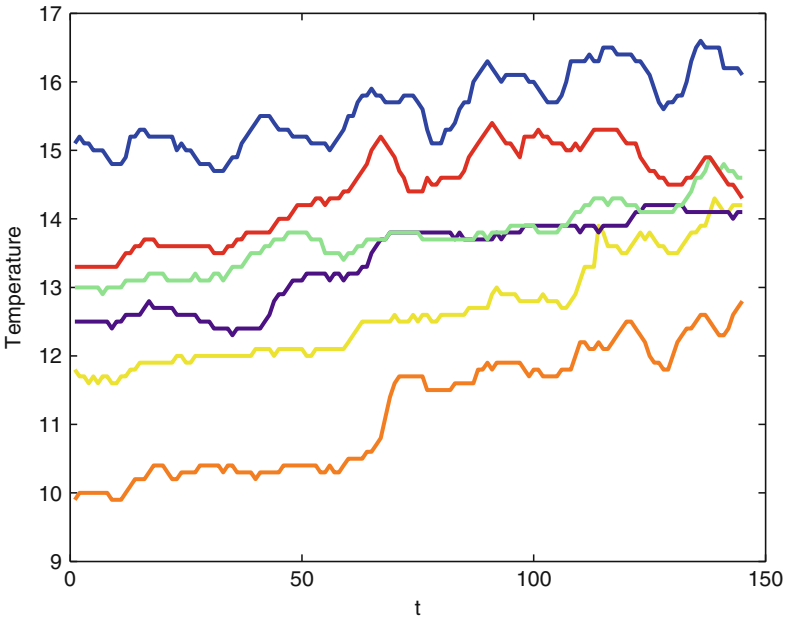


Fig. 1 Observed curve data for six locations along the Italian peninsula

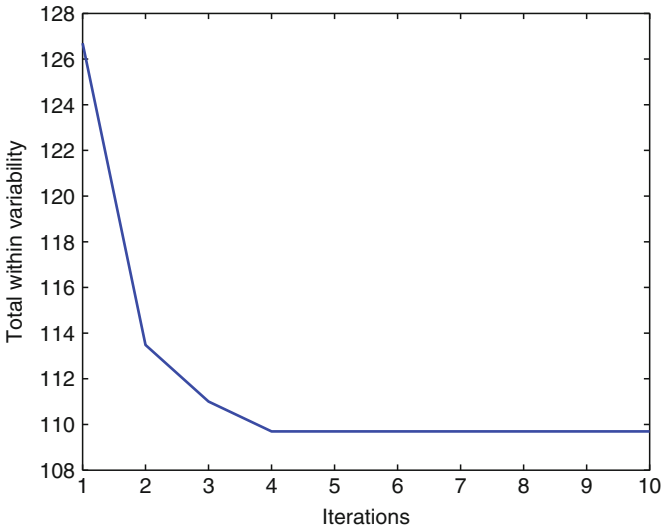


Fig. 2 Criterion value until the convergence

The objective of analyzing such kind of data is to find homogeneous spatial areas and, together, to find a localized functional model able to summarize their behaviors.

Since data are noisy and non periodic, B-spline basis functions appear to be an effective choice for getting the true functional form from the sample data. Especially we have used B-spline functions with an equi-spaced set of knots.

The first performed test, aims at evaluating the algorithm convergence. It is based on running the algorithm with several initializations and with a variable number of clusters C.

The convergence is proved by the attainment of a stable level of the optimized criterion which constitutes one of its local minima. In Fig. 2, the decrement of the criteria is shown for $C = 3$ clusters.

Further tests have been performed to analyze the clustering performances. By analyzing the decrement of the criterion (Fig. 3) for $C = 2, \ldots, 5$, we choose to get a partition of data into $C = 3$ clusters.

To initialize the clustering procedure, we have run a standard k-means algorithm on the spatial locations of the observed data. This is to get a partitioning of data into spatially contiguous regions.

Since the proposed method, detects the prototypes of each cluster starting from a regular spatial grid, we have to choose the spatial distance. Such choice depends from the detail level in the spatial location of the representative models, however, a finer grid impacts on the computational requirements since an higher number of functional kriging has to be computed. For our experiments, the cell distance has been set to 10 km.

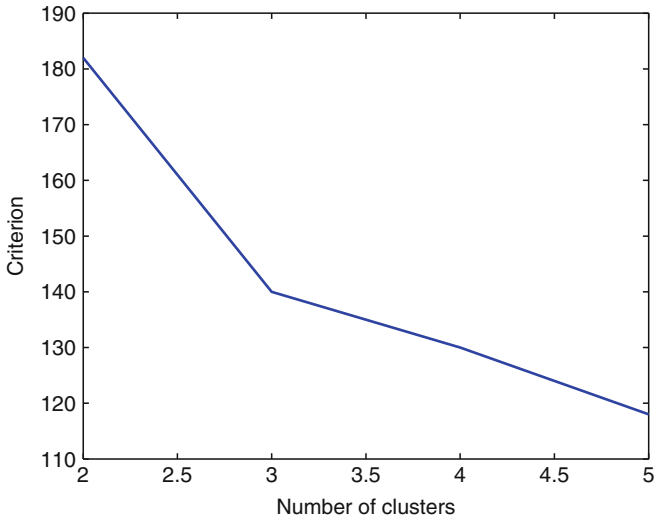


Fig. 3 Value of the optimized criterion for $C = 2, \ldots, 5$

The obtained cluster contains respectively $12, 9, 5$ elements, and have prototypes, as can be seen from the Fig. 4, located in three different zones of the coast: Sorrento, Francavilla Al Mare and Alassio Table 1.

The different shape of each prototype Fig. 5 shows three different behaviors of the involved regions. Looking at the clustering structure we can observe that:

- In the first cluster, according to the obtained functional parameters $\lambda_i, i = 1, \ldots, 12$, the greatest contribute to the prototype estimation corresponds to 0.63. This functional parameter corresponds to Salerno;
- In the second cluster, the greatest functional parameter that contributes to the prototype estimation corresponds to 0.49, this functional parameter corresponds to Ortona;
- In the third cluster, two functional parameters influence the prototype estimation. These are $\lambda_2 = 0.3, \lambda_5 = 0,29$ correspondent to Genova and La Spezia.

The obtained partition is such to divide the italian coast into three homogeneous zones representing three macro area of the sea respectively: Tirreno sea, Adriatico sea and Ligure sea.

6 Conclusion and Future Research

In the present paper we have introduced a new clustering strategy for spatio-functional data which is ables to discover representative functions for each cluster.

In the future research, further attention should be given to the development of more robust procedure for prototype estimation and of a criterion for outliers

Fig. 4 Locations of the three prototypes for each cluster (*red points*) and the network distribution of the station (*yellow points*)

Table 1 Locations of the prototypes

Prototype	Latitude	Longitude
Sorrento	40°37′53.98″	14°21′52.69″
Francavilla Al Mare	42°25′14.26″	14°17′15.91″
Alassio	44°00′10.58″	8°09′34.32″

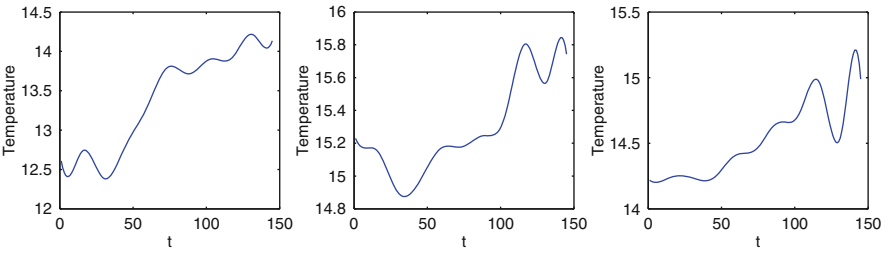


Fig. 5 Prototypes for each cluster

discover. Moreover tests on a wider set of data will be performed to evaluate the effectiveness of the procedure in several applicative contexts.

References

Abraham, C., Corillon, P., Matnzer-Löber, E., & Molinari, N. (2005). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics, 30*, 581–595.

Blekas, K., Nikou, C., Galatsanos, N., & Tsekos, N. V. (2007). Curve clustering with spatial constraints for analysis of spatiotemporal data. *In Proceedings of the 19th IEEE international Conference on Tools with Artificial intelligence* (Vol. 1, pp. 529–535). October 29–31. ICTAI. IEEE Computer Society, Washington, DC.

Delicado, P., Giraldo, R., & Mateu, J. (2007). *Geostatistics for functional data: An ordinary kriging approach.* Technical Report, http://hdl.handle.net/2117/1099, Universitat Politecnica de Catalunya.

Delicado, P., Giraldo, R., Comas, C., & Mateu, J. (2009). *Statistics for spatial functional data: some recent contributions*. Environmetrics, *21*(3–4), 224–239.

Diday, E. (1971). La Mthode des nues dynamiques. *Review on Statistical Application, XXX*(2), 19–34.

Heckman, N., & Zamar, R. (2000). Comparing the shapes of regression functions. *Biometrika, 87*, 135–144.

James, G., & Sugar, C. (2005). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association, 98*, 397–408.

Ramsay, J. O. (2008). *Fda problems that I like to talk about.* Personal communication.

Ramsay, J. E., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.) Berlin, Heidelberg, New York: Springer.

Romano, E. (2006). *Dynamical curves clustering with free knots spline estimation.* PHD Thesis. University of Federico II, Naples.

Use of Mixture Models in Multiple Hypothesis Testing with Applications in Bioinformatics

Geoffrey J. McLachlan and Leesa Wockner

Abstract There are many important problems these days where consideration has to be given to carrying out hundreds or even thousands of hypothesis testing problems at the same time. For example, in forming classifiers on the basis of high-dimensional data, the aim might be to select a small subset of useful variables for the prediction problem at hand. In the field of bioinformatics, there are many examples where a large number of hypotheses have to be tested simultaneously. For example, a common problem is the detection of genes that are differentially expressed in a given number of classes. The problem of testing many hypotheses at the same time can be expressed in a two-component mixture framework, using an empirical Bayes approach; see, for example, Efron (2004). In this framework, we present further results as part of an ongoing investigation into the approach of McLachlan et al. (2006) on the adoption of normal mixture models to provide a parametric approach to the estimation of the so-called local false discovery rate. The latter can be viewed as the posterior probability that a given null hypothesis does hold. With this approach, not only can the global false discovery rate be controlled, but also the implied probability of a false negative can be assessed. The methodology is demonstrated on some problems in bioinformatics.

1 Introduction

The analysis of very large data sets presents many challenges. One is the need to carry out the simultaneous testing of hundreds or possibly thousands of statistical hypotheses. In situations with many variables, an initial step in many analyses is to reduce the dimension of the problem by selecting a subset of useful variables by consideration of the variables considered separately. For example, in a situation

G.J. McLachlan (✉)
Department of Mathematics and Institute for Molecular Bioscience,
University of Queensland, Australia
e-mail: gjm@maths.uq.edu.au

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_18,

where the problem is to cluster the data at hand, the relevance of a variable might be assessed in terms of its effectiveness in revealing some group structure in the data. After formulating a null and an alternative hypothesis for each variable to assess its relevance for the problem at hand, we subsequently obtain a P-value for each variable. The question then arises on how to select a useful subset of variables from the many P-values in the case of a high-dimensional data set.

A problem as described above can occur in many modern scientific studies. To present our methodology, we shall focus on the problem in bioinformatics arising in the analysis of microarray experiments, known as the detection of differential expression. With this problem, the aim is to determine which of several thousands genes are differentially expressed between a number of k different classes $C_1, \ldots, C_k$. In the case of $k = 2$, Class C_1 might refer to some women who have a good prognosis following the diagnosis of a disease (such as breast cancer) while the other Class C_2 refers to those women who have poor prognosis, corresponding to the occurrence of distant metastases within 5 years.

In classic situations involving only one single hypothesis test, the aim is to control the probability of making a Type I error; that is, the probability of making a false positive. In situations where multiple (N) hypotheses are under test, one can use the Bonferroni method to control the probability that at least one false positive error will be made. In Table 1, we have listed the possible outcomes from N hypothesis tests.

However, in the current context where N is a very large number, controlling the family wise error rate (that is, the probability that $N_{01} \geq 1$) is too strict and will lead to missed findings. In our example of the detection of genes differentially expressed between the classes, the goal is to identify as many genes with significant differences as possible, while incurring a relatively low proportion of false positives.

In a seminal paper, Benjamini and Hochberg (1995) introduced a new multiple hypothesis testing error measure called the false discovery rate (FDR), which they defined as

$$\text{FDR} = E\{\frac{N_{01}}{N_r \vee 1}\}, \tag{1}$$

where $N_r \vee 1 = \max(N_r, 1)$. The effect of $N_r \vee 1$ in the denominator of the expectation in (1) is to set $N_{01}/N_r = 0$ when $N_r = 0$. They proposed an FDR-controlling step-up test procedure for independent P-values associated with the N hypotheses.

Other error rates in addition to the FDR are of interest in practice, such as the false non-discovery rate (FNDR) and the false negative rate (FNR), as given empirically in Table 1 by the ratios, $N_{10}/(N - N_r)$ and N_{10}/N_1, respectively. As the FNDR is nearly always quite small since N_{00} is usually much larger than N_{10}, the FNR is generally more informative.

In this paper, we concentrate on a parametric approach to the handling of the P-values to provide a procedure that not only can be used to control the FDR, but also allows the implied FNR to be estimated. Previously, Allison et al. (2002) had considered mixture modelling of the P-values directly in terms of a mixture of beta distributions with the uniform (0,1) distribution (a special form of a beta distribution) as the null component.

Table 1 Possible outcomes from N hypothesis tests

	Accept null	Reject null	Total
Null true	N_{00}	N_{01}	N_0
Non-true	N_{10}	N_{11}	N_1
Total	$N - N_r$	N_r	N

We adopt the parametric approach of McLachlan et al. (2006) that transforms the P_j values via the probit transformation to z-scores. Suppose in the present context of the detection of differential expression P_j denotes the P-value for the test of the null hypothesis

$$H_j : j\text{th gene is not differentially expressed.}$$

Then the z_j-score is given by

$$z_j = \Phi^{-1}(1 - P_j),$$

where Φ denotes the (cumulative) normal distribution function. This transformation is defined so that large positive values of the z_j-score suggest departures from the null hypothesis. Here the P_j values ($j = 1, \ldots, N$) constitute the input for this parametric approach. We do not consider how the P_j values are computed in the first instance. For example, they could be calculated on the basis of the classical t- or F-statistics, depending on whether there are two or multiple classes. Alternatively, the P_j might be calculated via a permutation method.

2 Modelling of Z-Scores

The density of the z_j-score can be modelled by a two-component mixture model as formulated in Lee et al. (2000) and Efron et al. (2001). We let G denote the population of genes under consideration. It can be decomposed into two groups G_0 and G_1, where G_0 is the group of genes that are not differentially expressed, and G_1 is the complement of G_0; that is, G_1 contains the genes that are differentially expressed. We let π_i denote the prior probability of a gene belonging to G_i ($i = 0, 1$), and we denote the density of z_j in G_i by $f_i(z_j)$. The unconditional density of Z_j is then given by the two-component mixture model,

$$f(z_j) = \pi_0 f_0(z_j) + \pi_1 f_1(z_j).$$

Using Bayes Theorem, the posterior probability that the jth gene is not differentially expressed (that is, belongs to G_0) is given by

$$\tau_0(z_j) = \pi_0 f_0(z_j)/f(z_j) \quad (j = 1, \ldots, N). \tag{2}$$

In this framework, the gene-specific posterior probabilities provide the basis for optimal statistical inference about differential expression. The posterior probability $\tau_0(z_j)$ has been termed the local false discovery rate (local FDR) by Efron and Tibshirani (2002). It quantifies the gene-specific evidence for each gene. As noted by Efron (2004), it can be viewed as an empirical Bayes version of the Benjamini-Hochberg (1995) methodology, using densities rather than tail areas.

It can be seen from (2) that in order to use this posterior probability of non-differential expression in practice, we need to be able to estimate π_0, the mixture density $f(z_j)$, and the null density $f_0(z_j)$, or equivalently, the ratio of densities $f_0(z_j)/f(z_j)$. Efron et al. (2004) has developed a simple empirical Bayes approach to this problem with minimal assumptions. We focus on a fully parametric approach using mixtures of normal densities. If the assumptions under which the P-values have been calculated hold, then the null density of Z_j is given by the standard normal density; that is,

$$f_0(z_j) = \phi(z_j; \mu_0, \sigma_0^2),$$

where $\mu_0 = 0$ and $\sigma_0^2 = 1$. This is known as the theoretical null distribution to distinguish it from the "empirical" null (as termed by Efron (2004)) in situations where the assumptions breakdown. The density $f_1(z_j)$ of z_j under the alternative hypothesis is approximated by a single normal density,

$$f_1(z_j) = \phi(z_j; \mu_1, \sigma_1^2).$$

In practice, the differentially expressed genes have varying values for the differences between their class means, and so it is somewhat surprising that for the data sets that we have analysed, a single normal distribution has sufficed to model the density of the z-scores for the non-null genes (genes that are differentially expressed). As shown by McLachlan et al. (2006), the two-component normal mixture model

$$f(z_j) = \pi_0\phi(z_j; 0, 1) + \pi_1\phi(z_j; \mu_1, \sigma_1^2)$$

can be fitted very quickly via the EM algorithm, as in their program called EMMIX-FDR.

The genes can be ranked on the basis of the estimated posterior probabilities $\tau_0(z_j)$, and we can select all genes with

$$\hat{\tau}_0(z_j) \leq c_o \tag{3}$$

to be differentially expressed. McLachlan et al. (2006) have shown how estimates of the implied rates, including the FDR and FNR, can be formed in terms of the $\tau_0(z_j)$ for a specified threshold c_o. In particular, an estimate of the FDR is given by

$$\widehat{FDR} = \sum_{j=1}^{N} \frac{\hat{\tau}_0(z_j) I_{[0,c_0]}(\hat{\tau}_0(z_j))}{N_r}, \tag{4}$$

where $I_A(x)$ is the indicator function, which is one if $x \in A$ and is zero otherwise.

3 Example: Breast Cancer Data

To illustrate the application of this parametric approach to multiple hypothesis testing, we consider the detection of differentially expressed genes for some data from the study of Hedenfalk et al. (2001), which examined gene expressions in breast cancer tissues from women who were carriers of the hereditary BRCA1 or BRCA2 gene mutations, predisposing to breast cancer. The data set comprised the measurement of $N = 3,226$ genes using cDNA arrays, for $n_1 = 7$ BRCA1 tumours and $n_2 = 8$ BRCA2 tumours. We display the fitted mixture density in Fig. 1.

In Table 2, we have listed the FDR estimated from (4) for various levels of the threshold c_o in (3). It can be seen, for example, that if c_o is set equal to 0.1, then the estimated FDR is 0.06 and $N_r = 143$ genes would be declared to be differentially expressed.

4 Empirical Null

As pointed by Efron (2004), for some microarray data sets the normal scores do not appear to have the theoretical null distribution, which is the standard normal. In this case, Efron has considered the estimation of the actual null distribution called the empirical null as distinct from the theoretical null.

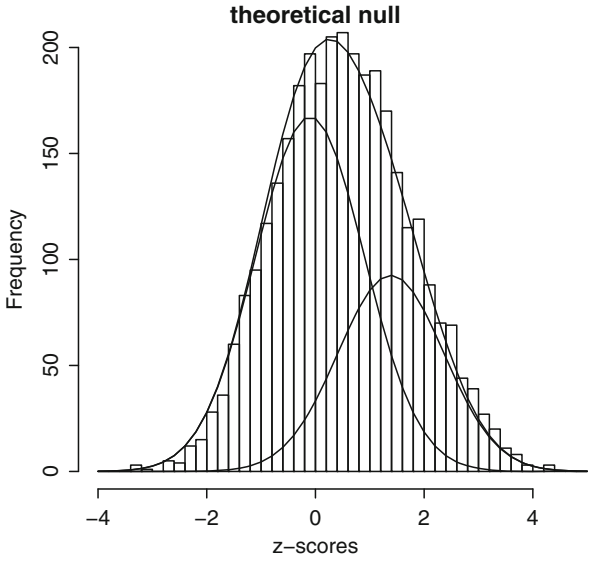


Fig. 1 Breast cancer data: plot of fitted two-component normal mixture model with theoretical $N(0, 1)$ null and non-null components (weighted respectively by $\hat{\pi}_0$ and $(1 - \hat{\pi}_0)$) imposed on histogram of z-scores

Table 2 Estimated FDR and other error rates for various levels of the threshold c_o applied to the posterior probability of nondifferential expression for the breast cancer data, where N_r is the number of selected genes (with theoretical null)

c_o	N_r	$\widehat{\text{FDR}}$	$\widehat{\text{FNDR}}$	$\widehat{\text{FNR}}$	$\widehat{\text{FPR}}$
0.1	143	0.06	0.32	0.88	0.004
0.2	338	0.11	0.28	0.73	0.02
0.3	539	0.16	0.25	0.60	0.04
0.4	743	0.21	0.22	0.48	0.08
0.5	976	0.27	0.18	0.37	0.13

If we adopt an empirical null in our parametric approach, we do not assume that the mean μ_0 and variance σ_0^2 of the null distribution are zero and one, respectively, but rather they are estimated in addition to the other parameters π_0, μ_1, and σ_1^2. One reason why the theoretical null distribution may not be appropriate is that the assumptions do not hold for the P-value to have a uniform distribution on the unit interval under the null hypothesis. Another reason is that the P-values are not independently distributed due to the expression profiles not being independent for all the genes.

5 Simulation Study

Allison et al. (2002) performed some simulations to investigate the effect of correlation among the genes on their results. They generated data for 10 tissue samples on 3,000 genes. Each gene profile was drawn from a 3,000-dimensional normal distribution with mean $\mu = 10$ and covariance matrix Σ. In order to mimic the idea that genes which are co-expressed would be correlated, but genes which are not co-expressed would not be correlated, Allison et al. (2002) split the 3,000 genes into six blocks of size 500. Within each block the correlation between the genes ranged over the three values of ρ: 0 (independence), 0.4 (moderate dependence), 0.8 (strong dependence). This resulted in a covariance matrix of the form

$$\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{B} \otimes \boldsymbol{I}_6, \tag{5}$$

where σ^2 is the common variance and where

$$\boldsymbol{B} = \mathbf{1}_{500}\mathbf{1}_{500}^T\rho + (1-\rho)\boldsymbol{I}_{500},$$

$\mathbf{1}_m$ is a vector of ones length m and $\boldsymbol{I}_m$ is the $m \times m$ identity matrix. Finally, for 20% of randomly selected genes (600 genes), a mean difference of Δ was added to the expression levels for the last five tissue samples. Allison et al. (2002) suggested that a value of $\rho = 0.4$ 'tended to produce higher correlations among gene expressions than were present in [their] actual example data set' contrary to previous opinions about the existence of correlations amongst gene expression levels.

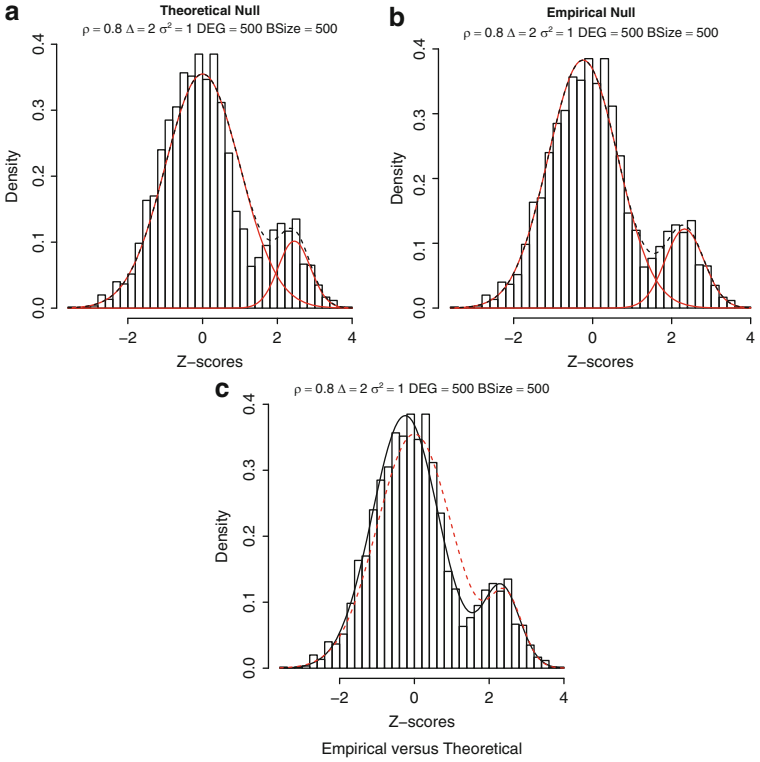


Fig. 2 A simulation study where $\pi_0 = 0.83$ and $\Delta = 2$. (a)–(c) Strong dependence (a) Overall Theoretical Null (*dashed*) with two (*weighted*) components (*solid*) $\hat{\pi}_0 = 0.88$ (b) Overall Empirical Null (*dashed*) with two (*weighted*) components (*solid*) $\hat{\pi}_0 = 0.85$ (c) An overlay of Theoretical (*dashed*) and Empirical (*solid*)

Following this example, we generated data for 10 tissue samples from a normal distribution with $\mu = 0$ and correlation matrix as in (5), with $\sigma^2 = 1$ and ρ defined as before. For 500 randomly selected genes (17%) a difference Δ of 1, 2 or 4 was added to the last five tissue samples.

It was demonstrated in the presence of strong correlation between the genes ($\rho = 0.8$) that the empirical null distribution led to a much better fit than with the theoretical null; see Fig. 2.

References

Allison, D. B., Gadbury, G. L., Heo, M., Fernández, J. R., Lee, C. K., Prolla, T. A., & Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis, 39*, 1–20.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, B57*, 289–300.

Efron, B. (2004). Selection and estimation for large-scale simultaneous inference. *Technical Report*. Department of Statistics, Stanford University, Stanford, CA, http://www-stat.stanford.edu/ brad/papers/Selection.pdf.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the Amercian Statistical Association, 99*, 96–104.

Efron, B., & Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology, 23*, 70–86.

Efron, B., Tibshirani, R., Storey, J. D., & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association, 96*, 1151–1160.

Hedenfalk, I., Duggan, D., Chen, Y. D., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P. et al. (2001). Gene-expression profiles in hereditary breast cancer. *The New England Journal of Medicine, 344*, 539–548.

Lee, M. M-T., Kuo, F. C., Whitmore, G. A., & Sklar, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences of the United States of America, 97*, 9834–9838.

McLachlan, G. J., Bean, R. W., & Ben-Tovim Jones, L. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics, 22*, 1608–1615.

Finding Groups in Ordinal Data: An Examination of Some Clustering Procedures

Marek Walesiak and Andrzej Dudek

Abstract The article evaluates, based on ordinal data simulated with cluster.Gen function of clusterSim package working in **R** environment, some cluster analysis procedures containing GDM distance for ordinal data (see Jajuga et al. 2003; Walesiak 1993, 2006), nine clustering methods and eight internal cluster quality indices for determining the number of clusters. Seventy two clustering procedures are evaluated based on simulated data originating from a variety of models. Models contain the known structure of clusters and differ in the number of true dimensions, the number of categories for each variable, the density and shape of clusters, the number of true clusters, the number of noisy variables. Each clustering result was compared with the known cluster structure from models applying (Hubert and Arabie 1985) corrected Rand index.

1 Introduction

Four basic scales are distinguished in the theory of measurement: nominal, ordinal, interval and ratio scale. Among these four scales of measurement the nominal is considered the lowest. It is followed by the ordinal, the interval, and the ratio one which is the highest. They were introduced by Stevens (1959).

Systematics of scales refers to transformations which retain relations of the respective scale. These results are well-known and presented e.g. in the paper Jajuga and Walesiak (2000), p. 106. Any strictly increasing functions are the only permissible transformations within the ordinal scale. The main characteristics of ordinal scale are summarised in Table 1.

M. Walesiak (✉)
Wrocław University of Economics, Nowowiejska 3, 58-500 Jelenia Góra, Poland
e-mail: marek.walesiak@ue.wroc.pl

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_19,

Table 1 Rules for ordinal scale of measurement

Scale	Basic empirical operations	Allowed mathematical transformations	Allowed arithmetic operations
Ordinal	Equal to, greater than, smaller than	Any strictly increasing functions	Counting of events (numbers of relations equal to, greater than, smaller than)

Source: Adapted from Stevens (1959), pp. 25, 27

2 Clustering Procedures for Ordinal Data

Major steps in cluster analysis procedure for ordinal data include (see e.g. Milligan 1996, pp. 341–343): the selection of objects and variables, the selection of a distance measure, the selection of clustering method, determining the number of clusters, cluster validation, describing and profiling clusters. Variable normalization step is omitted while performing comparisons with cluster analysis procedure for metric data. The purpose of normalization is to adjust the size and the relative weighting of input variables (see e.g. Milligan and Cooper 1988, p. 182). Normalization is used when variables are measured with metric data. Normalization is not necessary with regard to ordinal scale, because only the relations: equal to, greater than, smaller than are permitted with ordinal values.

The construction of distance measure for ordinal data should take these relations into account and should be based on relations between the two analyzed objects and the other objects (context distance measure). In statistical data analysis literature few distance measures for variables measured with ordinal data were suggested. Only GDM distance measure d_{ik} proposed by Walesiak (1993), pp. 44–45 satisfies ordinal scale conditions (see Table 1):

$$d_{ik} = \frac{1}{2} - \frac{\sum_{j=1}^{m} a_{ikj} b_{kij} + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} a_{ilj} b_{klj}}{\left[\sum_{j=1}^{m} \sum_{l=1}^{n} a_{ilj}^2 \sum_{j=1}^{m} \sum_{l=1}^{n} b_{klj}^2 \right]^{\frac{1}{2}}}, \tag{1}$$

$$a_{ipj}(b_{krj}) = \begin{cases} 1 \text{ if } x_{ij} > x_{pj}\,(x_{kj} > x_{rj}) \\ 0 \text{ if } x_{ij} = x_{pj}\,(x_{kj} = x_{rj}) \\ -1 \text{ if } x_{ij} < x_{pj}\,(x_{kj} < x_{rj}) \end{cases} \text{ for } p = k, l; r = i, l\,, \tag{2}$$

where: $i, k, l = 1, \ldots, n$ – the number of objects,
$j = 1, \ldots, m$ – the number of variables,
$x_{ij}\,(x_{kj}, x_{lj})$ – i-th (k-th, l-th) observation on the j-th variable.

Article Jajuga et al. (2003) discusses the properties of GDM distance measure.

Other proposals [e.g. Kendall distance measure (Kendall 1966, p. 181); Gordon distance (Gordon 1999, p. 19); Podani distance (Podani 1999)] imply the assumption that the ranks are measured with at least, the interval scale (when the

differences can be calculated). It is also worth mentioning the following argument, presented by Kaufman and Rousseeuw (1990), p. 30: "Therefore, most authors advice treating the ranks as interval-scaled and applying the usual formulas for obtaining dissimilarities (like the Euclidean or Manhattan distance)".

The selected clustering procedures included in the article are as follows:

1. GDM distance measure for ordinal data – GDM2 distance in clusterSim package.
2. The selected methods of cluster analysis (stats and cluster packages):
   - k-medoids (pam);
   - Seven hierarchical agglomerative algorithms: single link (single), complete link (complete), group average link (average), weighted average link (mcquitty), incremental sum of squares (ward), centroid (centroid), median (median). The Ward, centroid and median methods are easy to implement with distance matrix for only squared Euclidean distance. These methods could be used with any distance measure, however, the results would lack useful interpretation (see Anderberg 1973, pp. 141, 145);
   - Hierarchical divisive method by Macnaughton-Smith et al. (1964) – diana.
3. The selected internal cluster quality indices for determining clusters' number [all formulas and references for indices you can find in pdf files of clusterSim package (Walesiak and Dudek 2009)]: Davies-Bouldin – index.DB, Calinski-Harabasz – index.G1, Baker and Hubert – index.G2, Hubert and Levine – index.G3, gap – index.Gap, Hartigan – index.H, Krzanowski and Lai – index.KL, Silhouette – index.S.

For Davies-Bouldin, Calinski-Harabasz, gap, Hartigan, and Krzanowski and Lai indices medoids of clusters (representative objects of clusters) are used instead of centroids of clusters.

3 Simulation Experiment Characteristics

Data sets are generated in nine different scenarios (see Table 2). Models contain the known structure of clusters. Simulation models differ in the number of true dimensions (variables), the number of categories for each variable, the density and shape of clusters, the number of true clusters, the number of noisy (irrelevant) variables. The noisy variables are simulated independently, based on uniform distribution. Variations of noisy variables, in the generated data, are required to be similar to non-noisy ones (see Milligan 1985, Qiu and Joe 2006, p. 322).

The clusters in models presented in Table 2 contain continuous observations (metric data). Discretization process is performed on each variable in order to obtain ordinal data (see Walesiak and Dudek 2009). The number of categories k_j for categorical variable X_j determines the width of each class intervals $[\max_i\{x_{ij}\} - \min_i\{x_{ij}\}]/k_j$. Each class interval receives category $1, \ldots, k_j$ independently for

Table 2 Experimental factors for simulation models

m	*v*	*nk*	*cl*	*lo*	Centroid of clusters	Covariance matrix $\sum$	*ks*
1	2	4, 6	3	60, 30, 30	(0; 0), (1.5; 7), (3; 14)	$\sigma_{jj} = 1, \quad \sigma_{jl} = -0.9$	1
2	3	7	3	45	(1.5; 6, – 3), (3; 12; –6) (4.5; 18; –9)	$\sigma_{jj} = 1\ (1 \leq j \leq 3)$, $\sigma_{12} = \sigma_{13} = -0.9, \quad \sigma_{23} = 0.9$	1
3	2	5, 7	5	50, 20, 25, 25, 20	(5; 5), (–3; 3), (3; –3), (0; 0), (–5; –5)	$\sigma_{jj} = 1, \quad \sigma_{jl} = 0.9$	2
4	3	5, 7, 5	5	25	(5; 5; 5), (–3; 3; –3), (3; –3; 3), (0; 0; 0), (–5; –5; –5)	$\sigma_{jj} = 1\ (1 \leq j \leq 3)$, $\sigma_{jl} = 0.9\ (1 \leq j \neq l \leq 3)$	2
5	2	5	5	20, 45, 15, 25, 35	(0; 0), (0; 10), (5; 5), (10; 0), (10; 10)	$\sigma_{jj} = 1, \quad \sigma_{jl} = 0$	3
6	2	6, 8	4	35	(–4; 5), (5; 14), (14; 5), (5; –4)	$\sigma_{jj} = 1, \quad \sigma_{jl} = 0$	3
7	3	6	4	25, 25, 40, 30	(–4; 5; –4), (5; 14; 5), (14; 5; 14), (5; –4; 5),	a	4
8	3	5, 6, 7	5	35, 25, 25, 20, 20	(5; 5; 5), (–3; 3; –3), (3; –3; 3), (0; 0; 0), (–5; –5; –5)	b	4
9	2	7	3	40	(0; 4), (4; 8), (8; 12)	c	4

m – model, *v* – number of variables, *nk* – number of categories (one number means the same number of categories for each variable), *cl* – number of clusters, *lo* – number of objects in each cluster (one number means that clusters contain the same number of objects), *ks* – shape of clusters (1 – elongated, 2 – elongated and not well separated, 3 – normal, 4 – different for each cluster),

a: $\sum_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \sum_2 = \begin{bmatrix} 1 & -0.9 & -0.9 \\ -0.9 & 1 & 0.9 \\ -0.9 & 0.9 & 1 \end{bmatrix}, \sum_3 = \begin{bmatrix} 1 & 0.9 & 0.9 \\ 0.9 & 1 & 0.9 \\ 0.9 & 0.9 & 1 \end{bmatrix}, \sum_4 = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & 2 \\ 2 & 2 & 3 \end{bmatrix};$

b: $\sum_1 = \begin{bmatrix} 1 & -0.9 & -0.9 \\ -0.9 & 1 & 0.9 \\ -0.9 & 0.9 & 1 \end{bmatrix}, \sum_2 = \begin{bmatrix} 0.5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \sum_3 = \begin{bmatrix} 1 & 0.9 & 0.9 \\ 0.9 & 1 & 0.9 \\ 0.9 & 0.9 & 1 \end{bmatrix},$

$\sum_4 = \begin{bmatrix} 1 & 0.6 & 0.6 \\ 0.6 & 1 & 0.6 \\ 0.6 & 0.6 & 1 \end{bmatrix}, \sum_5 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix};$

c: $\sum_1 = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}, \sum_2 = \begin{bmatrix} 1.5 & 0 \\ 0 & 1.5 \end{bmatrix}, \sum_3 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}.$

Source: authors' compilation with clusterSim package (see Walesiak and Dudek 2009)

each variable and the actual value of variable x_{ij} is replaced by these categories. The number of categories may be different for each variable. The example of dicretization process is shown in Fig. 1.

The next step was to perform one out of seventy two clustering procedures (containing GDM distance for ordinal data, nine clustering methods and eight internal cluster quality indices for determining the number of clusters) with each model. The

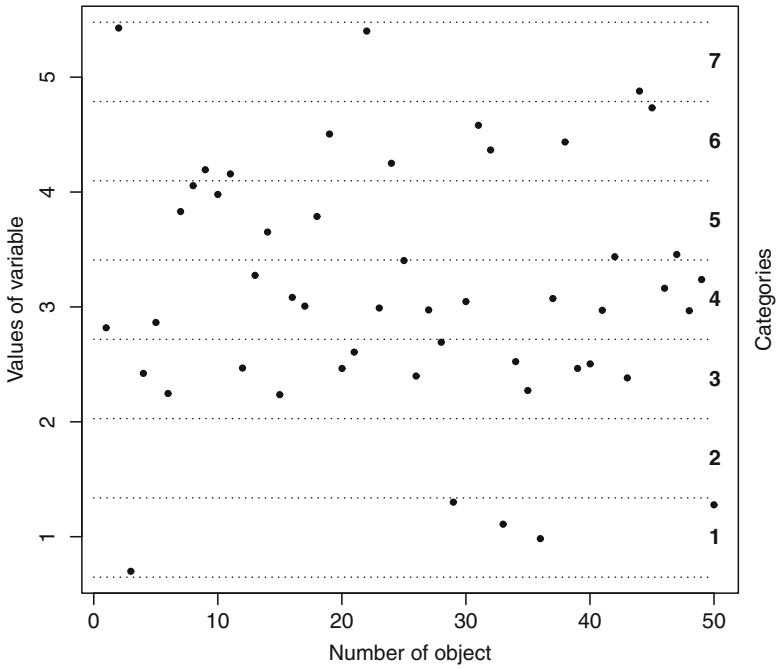


Fig. 1 The example of discretization process
Source: authors' compilation

analysis referred only to clustering results from 2 to 10 clusters. Next each clustering result was compared with the cluster structure known from models applying Hubert and Arabie (1985) corrected Rand index. The maximum value of corrected Rand index is 1 for identical partitions and its expected value is zero when the partitions are selected at random. Fifty realizations were generated from each setting.

4 Discussion on Simulation Results

In Table 3 nine clustering methods are ranked, based on adjusted Rand index mean values for nine models and eight internal cluster quality indices (with 50 simulations).

The following conclusions can be drawn from the results presented in Table 3:

– Group average method is definitely the best, while single link method is the worst for clustering ordinal data,
– Ward method ensures better results in clustering ordinal data with noisy variables.

Table 4 presents internal quality indices of clustering results ranking based on adjusted Rand index mean values for nine models and nine clustering methods (with 50 simulations).

Table 3 Clustering methods ranking based on adjusted Rand index mean values

Method	Mean		Shape of clusters								No. of noisy variables					
			1		2		3		4		0		2		4	
Average	0.545	1	0.514	1	0.509	2	0.494	1	0.625	1	0.739	1	0.508	1	0.388	1
Ward	0.512	2	0.473	3	0.479	3	0.465	2	0.591	3	0.680	7	0.482	2	0.373	2
Mcquitty	0.506	3	0.450	4	0.473	4	0.445	3	0.606	2	0.706	3	0.463	3	0.350	4
Diana	0.499	4	0.477	2	0.532	1	0.388	6	0.565	5	0.704	4	0.428	6	0.364	3
Complete	0.484	5	0.433	5	0.466	5	0.418	5	0.573	4	0.700	5	0.436	5	0.315	5
Pam	0.465	6	0.415	6	0.446	6	0.425	4	0.539	6	0.664	8	0.422	7	0.310	6
Centroid	0.408	7	0.384	7	0.362	7	0.370	7	0.479	8	0.721	2	0.451	4	0.051	8
Median	0.402	8	0.343	8	0.362	8	0.341	8	0.510	7	0.690	6	0.381	8	0.136	7
Single	0.312	9	0.324	9	0.238	9	0.256	9	0.390	9	0.613	9	0.291	9	0.032	9

Shape of clusters: 1 – elongated, 2 – elongated and not well separated, 3 – normal, 4 – different for each cluster

Table 4 Internal quality indices of clustering results ranking based on adjusted Rand index mean values

Method	Mean		Shape of clusters								No. of noisy variables					
			1		2		3		4		0		2		4	
KL	0.472	1	0.424	2	0.432	1	0.440	1	0.553	1	0.722	1	0.442	1	0.254	2
G1	0.430	2	0.422	3	0.406	4	0.352	5	0.503	3	0.616	4	0.423	2	0.250	3
Gap	0.414	3	0.440	1	0.323	8	0.341	6	0.505	2	0.687	2	0.346	7	0.208	8
G3	0.408	4	0.359	6	0.421	2	0.353	4	0.469	6	0.559	8	0.408	3	0.257	1
S	0.404	5	0.381	4	0.373	5	0.339	7	0.482	4	0.585	6	0.399	4	0.226	5
H	0.397	6	0.368	5	0.370	6	0.327	8	0.479	5	0.594	5	0.361	6	0.234	4
G2	0.391	7	0.313	8	0.406	3	0.358	3	0.456	7	0.583	7	0.373	5	0.218	6
DB	0.391	8	0.343	7	0.362	7	0.373	2	0.454	8	0.628	3	0.337	8	0.208	7

KL – Krzanowski and Lai, G1 – Caliński-Harabasz, Gap – gap, G3 – Hubert and Levine, S – Silhouette, H – Hartigan, G2 – Baker and Hubert, DB – Davies-Bouldin

Based on the results in Table 4 the following conclusions can be drawn:

– Krzanowski and Lai and Calinski and Harabasz indices present the best results in searching for optimal number of clusters in ordinal data,
– gap and Davies-Bouldin indices definitely show worse results in searching for optimal number of clusters in ordinal data containing noisy variables.

Table 5 presents the ranking of seventy two clustering procedures based on adjusted Rand index mean values for nine models and 50 simulations.

With reference to the aggregated results of simulations illustrated in Table 5 the following conclusions can be made:

– Clustering with group average link algorithm turns out to be the most efficient way for the simulation experiment, while applying Krzanowski and Lai index. This method, combined with Gap, Hartigan, Calinski-Harabasz and Davies-Bouldin indices, was ranked respectively at the fourth, sixth, seventh and ninth position,

Table 5 Clustering procedures ranking based on adjusted Rand index mean values (the selected results)

Rank	Method	Mean	Index	Shape of clusters								No. of noisy variables					
				1		2		3		4		0		2		4	
1	average	0.623	KL	0.553	7	0.577	1	0.608	1	0.710	1	0.853	3	0.590	1	0.426	1
2	ward	0.610	KL	0.537	9	0.550	5	0.596	2	0.708	2	0.852	4	0.571	2	0.407	4
3	ward	0.578	Gap	0.648	2	0.447	39	0.495	7	0.673	3	0.857	2	0.502	11	0.375	14
4	average	0.573	Gap	0.649	1	0.440	46	0.496	6	0.662	4	0.883	1	0.481	18	0.354	24
5	mcquitty	0.565	KL	0.488	16	0.528	8	0.533	4	0.662	5	0.801	9	0.512	9	0.381	13
6	average	0.564	H	0.556	6	0.531	7	0.471	12	0.654	6	0.726	19	0.544	3	0.423	2
7	average	0.558	G1	0.565	4	0.518	10	0.476	11	0.634	10	0.735	16	0.543	4	0.395	8
8	pam	0.553	KL	0.476	21	0.508	13	0.534	3	0.647	7	0.845	5	0.478	19	0.336	30
9	average	0.538	DB	0.486	17	0.502	16	0.530	5	0.601	18	0.772	14	0.474	20	0.367	18
10	diana	0.535	KL	0.466	23	0.571	3	0.457	16	0.609	16	0.780	12	0.458	28	0.367	17
–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
68	median	0.334	DB	0.267	69	0.288	65	0.313	60	0.425	66	0.678	35	0.266	68	0.059	61
69	single	0.292	S	0.302	67	0.247	69	0.228	70	0.358	69	0.618	60	0.250	69	0.008	66
70	single	0.269	DB	0.253	72	0.200	70	0.246	69	0.342	70	0.614	61	0.182	70	0.012	65
71	single	0.243	Gap	0.259	70	0.132	72	0.205	71	0.331	71	0.571	71	0.150	71	0.007	67
72	single	0.235	H	0.254	71	0.137	71	0.181	72	0.322	72	0.551	72	0.146	72	0.007	69

– The second and the third positions were taken by Ward method, along with applying Krzanowski and Lai and Gap indices,
– The single-link algorithm, combined with Hartigan, Gap and Davies-Bouldin indices, is the least efficient method for ordinal data clustering.

5 Limitations

In our analysis the random generation of data set comes from multivariate normal distribution in which clusters' locations and the homogeneity of shapes are defined by means (centroids) and covariance matrices (distortion of objects). Such approach is typical for many other simulation studies, presented e.g. in papers Soffritti (2003), Tibshirani and Walther (2005), Tibshirani et al. (2001). The infinite number of cluster shapes for any number of dimensions becomes the main problem regarding data generation with known cluster structure. It seems substantiated to consider other distributions and copula functions in data generation process for data with non-standard cluster shapes. This task poses substantial difficulties, especially in case of ordinal data.

In our simulation study we do not take into account such methods like as spectral clustering for ordinal data and non-distance based methods (e.g. Latent Class Analysis for ordinal data).

References

Anderberg, M. R. (1973). *Cluster analysis for applications.* New York, San Francisco, London: Academic Press.

Gordon, A. D. (1999). *Classification.* London: Chapman & Hall/CRC.

Hubert, L. J., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193–218.

Jajuga, K., & Walesiak, M. (2000). Standardisation of data set under different measurement scales. In R. Decker & W. Gaul (Eds.), *Classification and information processing at the turn of the millennium* (pp. 105–112). Berlin, Heidelberg: Springer-Verlag.

Jajuga, K., Walesiak, M., & Bąk, A. (2003). On the general distance measure. In M. Schwaiger & O. Opitz (Eds.), *Exploratory data analysis in empirical research* (pp. 104–109). Berlin, Heidelberg: Springer-Verlag.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley (second editon: 2005).

Kendall, M. G. (1966). Discrimination and classification. In P. R. Krishnaiah (Ed.), *Multivariate analysis I* (pp. 165–185). New York: Academic Press.

Macnaughton-Smith, P., Williams, W. T., Dale, M. B., & Mockett, L. G. (1964). Dissimilarity analysis: A new technique of hierarchical sub-division. *Nature, 202*, 1034–1035.

Milligan, G. W. (1985). An algorithm for generating artificial test clusters. *Psychometrika, 50*(1), 123–127.

Milligan, G. W. (1996). Clustering validation: results and implications for applied analyses. In P. Arabie, L. J. Hubert & G. de Soete (Eds.), *Clustering and classification* (pp. 341–375). Singapore: World Scientific.

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification, 5*(2), 181–204.

Podani, J. (1999). Extending Gower's general coefficient of similarity to ordinal characters. *Taxon, 48*, 331–340.

Qiu, W., & Joe, H. (2006). Generation of random clusters with specified degree of separation. *Journal of Classification, 23*(2), 315–334.

Soffritti, G. (2003). Identifying multiple cluster structures in a data matrix. *Communications in Statistics. Simulation and Computation, 32*(4), 1151–1177.

Stevens, S. S. (1959). Measurement, psychophysics and utility. In C. W. Churchman and P. Ratooch (Eds.), *Measurement. Definitions and theories* (pp. 18–63), New York: Wiley.

Tibshirani, R., & Walther, G. (2005). Cluster validation by predicting strength. *Journal of Computational and Graphical Statistics, 14*(3), 511–528.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society, ser. B, 63*(2), 411–423.

Walesiak, M. (1993). *Statystyczna analiza wielowymiarowa w badaniach marketingowych [Multivariate Statistical analysis in marketing research]*. Wrocław University of Economics, Research Papers no. 654.

Walesiak, M. (2006). *Uogólniona miara odległości w statystycznej analizie wielowymiarowej [The generalised distance measure in multivariate statistical analysis]*. Wrocław: Wydawnictwo AE.

Walesiak, M., & Dudek, A. (2009). *clusterSim package*, http://www.R-project.org/.

An Application of One-mode Three-way Overlapping Cluster Analysis

Satoru Yokoyama, Atsuho Nakayama, and Akinori Okada

Abstract In recent years, it is possible to more easily obtain multi-way data, and various analysis models for the data have been suggested by several researchers. However, only small number of studies have been done on the one-mode multi-way data analysis model. The authors suggested an overlapping cluster analysis model for one-mode three-way similarity data in our earlier study. In the present study, the authors make an attempt to improve the one of problems of the algorithm and validate the improved algorithm by analyzing the one-mode three-way similarity data which calculated from the panel data of meals. Then, the one-mode three-way result is compared with the results with the one-mode two-way similarity data analysis to show the usefulness of the one-mode three-way similarity data analysis.

1 Introduction

In recent years, it is easy to obtain multi-way data because of advancement in information technology. Various analysis models for multi-way data have been suggested by several researchers. For example, three-way data are treated in Gower and De Rooij (2003), Joly and Le Calvé (1995), and Nakayama (2005) All of them use three-way distances to express relationships among three objects. PARAFAC models (Harshman 1970, 1972) and Tucker 3 models (Tucker 1966) also treated three-way data. While these models can be applied to three-mode three-way data, they cannot be applied to one-mode three-way proximity (similarity/dissimilarity) data.

In Yokoyama et al. (2009), the authors suggested an overlapping cluster analysis model and an associated algorithm for one-mode three-way similarity data based on Shepard and Arabie (1979) and Arabie and Carroll (1980).

S. Yokoyama (✉)
Department of Business Administration, Faculty of Economics, Teikyo University, 359 Otsuka Hachiouji City, Tokyo, 192-0395, Japan
e-mail: satoru@main.teikyo-u.ac.jp

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_20,

In the present study, after explaining the one-mode two-way overlapping cluster analysis model, one-mode three-way model is introduced in brief. Then the improved algorithm of the one-mode three-way model is described to deal with one of the problems mentioned by Yokoyama et al. (2009). The improved algorithm is applied to the panel data of meals. After that, the result of the one-mode three-way analysis is compared with that of the one-mode two-way analysis. Finally, we summarize the present study, and mention the future plans.

2 Overlapping Cluster Analysis Models

Overlapping cluster analysis model was introduced by Shepard and Arabie (1979). This model is called ADCLUS. In this model, the similarity s_{ij} is represented as

$$s_{ij} \cong \sum_{r=1}^{R} w_r p_{ir} p_{jr} + c, \tag{1}$$

where w_r is the nonnegative weight of the r-th cluster and c is the additive constant. The p_{ir} is binary; if the object i belongs to cluster r, p_{ir} is 1, and otherwise it is 0.

An algorithm called MAPCLUS was introduced (Arabie and Carroll 1980), which is extended and improved from the algorithm of ADCLUS. In this algorithm, p_{ir} is initially considered to vary continuously. This algorithm uses a gradient approach to minimize a loss function that seeks to maximize the variance accounted for (VAF).

The loss function is the weighted sum of two terms; one is simply a normalized measure of the sum of squared error, the other consists of a penalty function to make the value of p_{ir} binary. w_r $(r = 1, \ldots, R)$ is estimated using multiple linear regressions. Actually, p_{ir} and w_r are computed using an alternative least squares approach. In addition, MAPCLUS algorithm are using other techniques to estimate p_{ir} and w_r; "polishing," "de novo iterations," and "combinatorial optimization." Polishing is applied to make p_{ir} binary. The de novo iteration is the re-estimating to eliminate negative weights. And combinatorial optimization is performed to improve the VAF further.

In Yokoyama et al. (2009), the authors suggested the one-mode three-way overlapping clustering model, and proposed an algorithm to fit that model to one-mode three-way similarity data. In this model, the one-mode three-way similarity s_{ijk} is represented as

$$s_{ijk} \cong \sum_{r=1}^{R} w_r p_{ir} p_{jr} p_{kr} + c. \tag{2}$$

The associated algorithm is basically extended from MAPCLUS to deal with one-mode three-way similarity data. The loss function fitting the model to the

similarity is

$$L_r = \alpha_r A_r + \beta_r B_r, \tag{3}$$

In (2), A_r is the term of normalized sum of squared error;

$$A_r = \frac{\sum_i^n \sum_{\substack{j \\ i>j}}^{n-1} \sum_{\substack{k \\ j>k}}^{n-2} (\delta_{ijk}^{(r)} - w_r p_{ir} p_{jr} p_{kr})^2}{\sum_i^n \sum_{\substack{j \\ i>j}}^{n-1} \sum_{\substack{k \\ j>k}}^{n-2} \delta_{ijk}^{(r)2} / {}_nC_3}, \tag{4}$$

and B_r is designed to force the pairwise products $p_{ir} p_{jr} p_{kr}$ to be 0 or 1;

$$B_r = \frac{\sum_i^n \sum_{\substack{j \\ i>j}}^{n-1} \sum_{\substack{k \\ j>k}}^{n-2} [(p_{ir} p_{jr} p_{kr} - 1) p_{ir} p_{jr} p_{kr}]^2}{\sum_i^n \sum_{\substack{j \\ i>j}}^{n-1} \sum_{\substack{k \\ j>k}}^{n-2} \left(p_{ir} p_{jr} p_{kr} - \frac{1}{{}_nC_3} \sum_i^n \sum_{\substack{j \\ i>j}}^{n-1} \sum_{\substack{k \\ j>k}}^{n-2} p_{ir} p_{jr} p_{kr} \right)^2}, \tag{5}$$

where $\delta_{ijk}^{(l)} = s_{ijk} - \sum_{\substack{r=1 \\ r \neq l}}^{R} w_r p_{ir} p_{jr} p_{jk} - c$ and $\alpha_r + \beta_r = 1$.

While the present algorithm is similar to MAPCLUS, there are several differences between two algorithms; the iterative procedure to estimate p_{ir} and w_r, combinatorial optimization, and the threshold values of the polishing. Details are shown in Yokoyama et al. (2009).

3 Improvement of the Algorithm

In Yokoyama et al. (2009), the authors analyzed one-mode three-way artificial data and joint purchase data, but they mentioned in Sect. 5 of Yokoyama et al. (2009) that the suggested algorithm needs two improvements. The first is the stability of the algorithm and the second is the negative weights. In the present paper, the authors execute the improvement of the algorithm to eliminate negative weights. This improvement is as follows: If objects belong to the cluster where the weight is negative, joint occurrence of these objects in the cluster has negative influence on the similarities. Thus, we temporarily take the absolute value of w_r and interchange the values of p_{ir} in the iterative procedures of estimating p_{ir} and w_r. Because the iterative procedure becomes endless, the number of the improvement is limited to 10 times in each iterative procedure, and the improvement satisfies that three or more values of p_{ir} should equal to 1 in cluster r.

To examine the effectiveness of this improvement, one-mode three-way similarity data shown in Table 2 are analyzed using two algorithms, one is with the improvement, another is without the improvement. The data are explained in Sect. 4 in detail. In these analyses, the data are analyzed using 10,000 different random initial values of p_{ir}, the maximum number of clusters is eight, and the minimum is

Table 1 The proportion of the frequencies of the results not having negative weights

Num. of Cluster	8	7	6	5	4	3
Without improvement	0.00%	0.00%	0.08%	0.56%	3.35%	27.89%
With improvement	0.01%	0.14%	0.46%	4.05%	20.75%	39.25%

three. The proportion of the frequencies of the results not having negative weights of these analyses are shown in Table 1.

The proportion had decreased by 5.5 points on the average due to the improvement. Especially, in four, five, and six clusters, the results not having negative weights increased drastically, and in seven and eight clusters, the results not having negative weights were able to be derived. However, negative weights still appear after the improvement, the fundamental improvement of the algorithm seems necessary.

4 An Application

In the present study, the panel data of meals are analyzed using with the improved algorithm. This data was provided by Data Analysis Competition 2008 (sponsored by Joint Association Study Group of Management Science). The data recorded the menus and foodstuffs for every meal during one year for about 200 families living in the Tokyo area, and consist of 150,000 meals including about 1,000 menus and about 2,000 foodstuffs. In the present study, we have paid attention to liquors which were drunk during each evening meal. In the evening meal, sometimes several liquors are served on the table, and sometimes the liquor is not served. For the present analysis we extracted meals where several liquors were served from the data, and one-mode three-way similarity data are calculated the frequencies of three liquors are jointly served.

The calculated data have more than 15 categories of liquors, but in the present analysis, nine categories, which have substantial frequency of being jointly served, are selected. The nine categories are; 1. Beer, 2. Burgundy, 3. White wine, 4. Champagne, 5. Whiskey, 6. Cocktails, 7. Sake, 8. Shochu (a clear liquor distilled from sweet potatoes, rice, etc.), and 9. Fruit liquor (incl. Chinese liquor). The derived one-mode three-way similarity data are shown in Table 2.

The similarity data were analyzed using the improved algorithm, the largest VAF for each number of clusters is regard to be the maximum VAF at that number of clusters. The resulting maximum VAF in eight through three clusters were 0.984, 0.984, 0.984, 0.978, 0.961, and 0.886. Because of the noticeable elbow of the VAF and the interpretation of the results, the six cluster result was chosen as the final solution shown in Table 3.

In the analysis, Beer belongs to Clusters 1, 2, 3, 4, and 6, Shochu belongs to Clusters 1, 3, 4, and 5. Especially, to Clusters 1, 3, and 4, Beer and Shochu belong together. Similarly, Beer and Burgundy belong together to Clusters 2 and 4, Beer

Table 2 One-mode three-way similarity data

		2.	3.	4.	5.	6.	7.	8.
1. Beer	3. White wine	9						
	4. Champagne	2	3					
	5. Whiskey	11	24	0				
	6. Cocktails	1	0	0	2			
	7. Sake	141	7	1	7	1		
	8. Shochu	77	198	3	24	12	103	
	9. Fruit liquor	1	1	1	1	1	10	12
2. Burgundy	4. Champagne		2					
	5. Whiskey		1	0				
	6. Cocktails		0	0	0			
	7. Sake		1	1	0	0		
	8. Shochu		3	1	0	1	37	
	9. Fruit liquor		1	1	0	0	1	1
3. White wine	5. Whiskey			0				
	6. Cocktails			0	0			
	7. Sake			1	0	0		
	8. Shochu			1	9	0	4	
	9. Fruit liquor			1	0	0	1	1
4. Champagne	6. Cocktails				0			
	7. Sake				0	0		
	8. Shochu				0	0	1	
	9. Fruit liquor				0	0	1	1
5. Whiskey	7. Sake					0		
	8. Shochu					2	2	
	9. Fruit liquor					0	0	0
6. Cocktails	8. Shochu						0	
	9. Fruit liquor						0	1
7. Sake	9. Fruit liquor							2

and White wine belong together to Clusters 1 and 6, Beer and Sake belong together to Clusters 2 and 3, Burgundy and Sake belong together to Clusters 2 and 5, and Sake and Shochu belong together to Clusters 3 and 5.

For the reason of the value of each weight, with Beer and Shochu, White wine is more frequently served than Sake or Burgundy is, because the weights are 0.990, 0.510, and 0.379 for Clusters 1, 3, and 4, respectively. Similarly, with Beer and Burgundy, Sake is more frequently served than Shochu. With Beer and White wine, Shochu is more frequently served than Whisky is. With Beer and Sake, Burgundy is more frequently served than Shochu is. With Burgundy and Sake, Beer is more frequently served than Shochu is. With Sake and Shochu, Beer is more frequently served than Burgundy is.

To compare with the one-mode three-way result, one-mode two-way similarity data are calculated the frequencies of two liquors are jointly served, and were

Table 3 The solution of one-mode three-way analysis

Category	Cluster					
	1	2	3	4	5	6
1. Beer	1	1	1	1	0	1
2. Burgundy	0	1	0	1	1	0
3. White wine	1	0	0	0	0	1
4. Champagne	0	0	0	0	0	0
5. Whiskey	0	0	0	0	0	1
6. Cocktails	0	0	0	0	0	0
7. Sake	0	1	1	0	1	0
8. Shochu	1	0	1	1	1	0
9. Fruit liquor	0	0	0	0	0	0
Weight	0.990	0.702	0.510	0.379	0.177	0.111

Table 4 One-mode two-way similarity data

	1.	2.	3.	4.	5.	6.	7.	8.
2. Burgundy	471							
3. White wine	286	31						
4. Champagne	23	3	6					
5. Whiskey	503	18	28	1				
6. Cocktails	73	3	2	1	3			
7. Sake	1,458	239	12	1	12	6		
8. Shochu	2,856	142	216	3	49	31	182	
9. Fruit liquor	60	1	1	1	1	1	28	24

Table 5 The solution of one-mode two-way analysis

Category	Cluster					
	1	2	3	4	5	6
1. Beer	1	1	1	1	1	0
2. Burgundy	0	0	0	1	0	1
3. White wine	0	0	0	0	1	0
4. Champagne	0	0	0	0	0	0
5. Whiskey	0	0	1	0	0	0
6. Cocktails	0	0	0	0	0	0
7. Sake	0	1	0	0	0	1
8. Shochu	1	0	0	0	1	1
9. Fruit liquor	0	0	0	0	0	0
Weight	0.908	0.505	0.172	0.160	0.084	0.063

analyzed by using one-mode two-way model. The derived one-mode two-way similarity data are shown in Table 4.

One-mode two-way similarity data were analyzed by using MAPCLUS. The resulting maximum VAF in eight through three clusters were 0.999, 0.999, 0.998, 0.992, 0.984, and 0.966. Same as the one-mode three-way analysis, the six cluster result was chosen as the final solution shown in Table 5.

In one-mode two-way analysis, Beer belongs to Clusters 1 through 5. In Clusters 1 through 4, Beer combines with Shochu, Sake, Whisky, and Burgundy. Especially, the weight of Clusters 1 and 2 is 0.908 and 0.505, and they are larger than the others. Therefore, with Beer, Shochu or Sake is more frequently served than Whisky or Burgundy is.

Here, we pay attention to Clusters 2 through 5 in one-mode three-way analysis, which Beer, Burgundy, Sake, or Shochu belongs to any of Clusters 2 through 5. These categories belong to Clusters 1, 2, and 4 in one-mode two-way analysis. From a viewpoint of the weight, the following six results are obtained:

With Beer and Burgundy, Sake is more frequently served than Shochu is,
with Beer and Sake, Burgundy is more frequently served than Shochu is,
with Beer and Shochu, Sake is more frequently served than Burgundy is,
with Burgundy and Sake, Beer is more frequently served than Shochu is,
with Burgundy and Shochu, Beer is more frequently served than Sake is, and
with Sake and Shochu, Beer is more frequently served than Burgundy is.

Here, from one-mode three-way similarity data in Table 2, we examine the frequencies of combinations among three categories in four categories, Beer, Burgundy, Sake and Shochu; Beer, Burgundy, and Sake is 141, Beer, Burgundy, and Shochu is 77, Beer, Sake, and Shochu is 103, and Burgundy, Sake, and Shochu is 37. Therefore, the six cases are substantiated by these frequencies completely. Similarly, the frequencies of combination of two categories from these four from Table 4 are substantiated the one-mode two-way results. However, in the combination between two categories by one-mode two-way analysis, we cannot grasp which combinations among three categories are frequently served, i.e., we can grasp the combination of Beer and Shochu have been frequently served by the one-mode two-way analysis, but we cannot grasp that if Sake or Burgundy is more frequently served with Beer and Shochu. Thus, one-mode three-way analysis can express the information which cannot be expressed by one-mode two-way analysis.

5 Discussion and Conclusion

In the present study, the authors introduced the one-mode three-way overlapping cluster analysis model and the associated algorithm in brief. Then the authors attempted to improve the one of problems of the algorithm, and compared with two algorithms. As the results shown in Table 1, the results not having negative weights had increased in all number of clusters by the improvement, that is, we can derive a result with smaller number of initial values. Thus the improved algorithm becomes more practical.

Moreover, one-mode three- and two-way similarity data, were analyzed by using one-mode three- and two-way overlapping cluster analysis models, respectively. Because these results were compared, the authors succeeded in indicating the necessity of one-mode three-way analysis.

In these analyses, the VAF were very large in the present analysis, and they are more than 0.9 for almost all cases. These are attributed to the frequencies of three and two liquors jointly served, so that these similarity data might be easy to divide into clusters by the analyses. It seems to be necessary to analyze various kinds of data, and examine the results carefully in future studies.

In the present study, the authors applied the overlapping cluster analysis model to one-mode three-way similarity data. In the future, it is necessary to analyze one-mode four (or more) -way, or two-mode four-way similarity data where individuals are added to one-mode three-way data. Thus, the present model should be extended to accommodate these data, and further improvements should be possible.

References

Arabie, P., & Carroll, J. D. (1980). MAPCLUS: A mathematical programming approach to fitting the ADCLUS model. *Psychometrika, 45*, 211–235.

Gower, J. C., & De Rooij, M. (2003). A comparison of the multidimensional scaling of triadic and dyadic distances. *Journal of Classification, 20*, 115–136.

Harshman, R. A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics, 16*, 1–84.

Harshman, R. A. (1972). Determination and proof of minimum uniqueness conditions for PARAFAC1. *UCLA working papers in phonetics, 22*, 111–117.

Joly, S., & Le Calvé, G. (1995). Three-way distances. *Journal of Classification, 12*, 191–205.

Nakayama, A. (2005). A multidimensional scaling model for three-way data analysis. *Behaviormetrika, 32*, 95–110.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review, 86*, 87–123.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika, 31*, 279–311.

Yokoyama, S., Nakayama, A., & Okada, A. (2009). One-mode three-way overlapping cluster analysis. *Computational Statistics, 24*, 165–179.

Evaluation of Clustering Results: The Trade-off Bias-Variability

Margarida G.M.S. Cardoso, Katti Faceli, and André C.P.L.F. de Carvalho

Abstract Clustering evaluation generally relies on some desirable properties of clustering solutions (partitions, in particular): the properties of clusters' compactness and separation, as well as the property of stability are often considered as indicators of clustering quality. In fact, since the real clustering is unknown (clustering being originated by an unsupervised process), one should focus on obtaining good enough partitions.

Clustering quality is, however, a difficult concept to put in practice. Furthermore, when aiming for clusters compactness and separation one does not necessarily meet the real clusters (e.g. Brun et al. 2007). Similarly, when focusing on the property of stability, one may find that solutions which are more stable but do not necessarily fit better the real solution (e.g. Cardoso et al. 2008).

In the present paper we consider clustering solution's reproducibility in other data sets drawn from the same source as an indicator of stability. We use a new cross-validation procedure and measure the agreement between clustering solutions obtained and the real partitions (real data sets from the UCI repository, Asunción and Newman 2007, are used). Next, we study the association between indicators of stability and agreement with the real partition. We conclude with a discussion of the trade-off bias-variability, which we believe is a relevant issue to investigate within unsupervised learning, clustering in particular.

1 Desirable Properties of a Clustering Solution

Ideally, clustering evaluation should take into account the degree of fit between the partition obtained (derived through cluster analysis) and the real or true partition. However, since the real partition is unknown (clustering is unsupervised),

M.G.M.S. Cardoso (✉)
Department of Quantitative Methods, ISCTE Business School, Av. das Forças Armadas 1649-026, Lisboa, Portugal
e-mail: margarida.cardoso@iscte.pt

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_21,

one should focus on the identification of a good enough partition, which exhibits some desirable properties.

The compactness and separation properties of a clustering structure define its quality: compactness measures the internal cohesion among objects within clusters; separation measures the isolation of clusters when compared to other clusters. Quality indices such as the Calinski and Harabasz (1974) and the Davies and Bouldin (1979) index are commonly used to evaluate these properties (e.g. Cardoso and Carvalho 2009).

Stability is also widely recognized as a desirable property of a clustering solution (e.g. Gordon 1999). A (stable) clustering solution should stay approximately the same when minor changes occur in the clustering procedure: alternative parameterizations of the clustering algorithm; introduction of noise in data; different clustering base variables; distinct data samples, etc.

2 Evaluating Stability

2.1 Indices of Agreement Between Partitions

Indices of agreement (IA) are frequently used to evaluate a clustering solution's stability, measuring the association between two partitions. They are generally based on data from a contingency table, built from the partitions (Π^K and Π^Q) being compared (see Cardoso and Carvalho 2009).

Several alternative IA can be found in the literature. The Rand index (Rand 1971), is, perhaps, the most popular of them. It quantifies the proportion of pairs classified in agreement by two partitions (1 indicating perfect agreement).

Hubert and Arabie (1985) studied the Rand index distribution under the hypothesis of random agreement, relying on the generalized hypergeometric model, and then suggested a modified Rand index. This Adjusted Rand index (AR) incorporates a threshold index value (expected index value under the null hypothesis) including a correction for agreement by chance:

Table 1 General cross-validation procedure

Step	Action	Output
1	Perform training-test sample split	Training and test samples
2	Cluster training sample	Clusters in the training sample
3	Build a classifier in training sample supervised by clusters' labels; use the classifier in the test sample.	Classes in the test sample
4	Cluster test sample	Clusters in the test sample
5	Obtain a contingency table between clusters and classes in the test sample and calculate indices.	Indices of agreement values, indicators of stability

$$AR\left(\Pi^K,\Pi^Q\right)=\frac{\sum_{k=1}^{K}\sum_{q=1}^{Q}\binom{n_{kq}}{2}-\sum_{k=1}^{K}\binom{n_{k.}}{2}\sum_{q=1}^{Q}\binom{n_{.q}}{2}\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum_{k=1}^{K}\binom{n_{k.}}{2}+\sum_{q=1}^{Q}\binom{n_{.q}}{2}\right]-\sum_{k=1}^{K}\binom{n_{k.}}{2}\sum_{q=1}^{Q}\binom{n_{.q}}{2}\Big/\binom{n}{2}}$$

The Variation of Information Index (VI) (Meila 2007) is based on Information Theory. This index measures how much of information is either lost or gained when objects are moved from one cluster to another (the smaller, the better, 0 indicating perfect agreement):

$$VI\left(\Pi^k,\Pi^Q\right)=H\left(\Pi^k\right)+H\left(\Pi^Q\right)-2I\left(\Pi^k,\Pi^Q\right)$$

where H indicates the entropy and I indicates the mutual information:

$$H\left(\Pi^K\right)=-\sum_{k=1}^{K}\frac{n_{k.}}{n}\log\left(\frac{n_{k.}}{n}\right)$$

$$I\left(\Pi^K,\Pi^Q\right)=\sum_{k=1}^{K}\sum_{q=1}^{Q}\frac{n_{kq}}{n}\log\left(n_{kq}\Big/\frac{n_{k.}n_{.q}}{n}\right)$$

2.2 Cross-validation

A common procedure for the evaluation of supervised analysis techniques is the use of cross-validation (Stone 1974). McIntyre and Blashfield (1980) import the concept of cross-validation from supervised to unsupervised analysis, clustering in particular. Breckenridge (1989) uses cross-validation and compares three alternative rules, Nearest Centroid, Quadratic Discriminant and Nearest Neighbour, to obtain classes in the test sample. This general cross-validation procedure is presented in Table 1.

Cross-validation may be replicated using alternative training-test splits, which provide several samples to evaluate the stability of a clustering solution (several IA values being considered; Tibshirani et al. 2005; Levine and Domany 2001; Dudoit and Fridlyand 2002; Law and Jain 2003; Lange et al. 2004; Cardoso 2007).

3 The Proposed Approach

When performing cross-validation, one needs to select an appropriate classifier to be trained using the clusters' labels in its training sample and able to obtain the desired classes within a test sample. According to Lange et al. (2004), "by selecting an inappropriate classifier, one can artificially increase the discrepancy between

solutions (...) the identification of optimal classifiers by analytical means seems unattainable. Therefore, we have to resort to potentially suboptimal classifiers in practical applications", (pp.1304–1305).

In this paper, the authors propose a new approach that overcomes the need for selecting a classifier when performing cross-validation. This approach relies on weighted samples and calculates the IA values based on the entire sample. Note that alternative approaches relying on sampling with replacement (e.g. Law and Jain 2003), also determine IA values based on the entire sample. However, they resort to a classifier to induce the clustering structure for all observations in the original sample.

The proposed approach – a weighted cross-validation procedure – enables to analyze the relationship between variability/stability and agreement with the real partition, when evaluating a clustering solution. It can be described as follows:

Cluster the original sample
Evaluate bias: obtain IA values between the partition being evaluated and the real one
For t = 1 to 20 do:
Draw a weighted sample from the original sample
Cluster the weighted sample
Obtain IA[t] values between the original and the weighted sample's partition
Evaluate variability: use 20 IA[t] values based on the weighted samples to obtain IA descriptive statistics

The well known K-Means (MacQueen 1967), and the maximum likelihood estimation of a finite mixture model (via a EM-Expectation-Maximization based algorithm, (Dempster et al. 1977; Vermunt and Magidson 2002), are used to obtain alternative clustering solutions.

In order to produce the weighted sample, drawn from the original sample, a unit weight is (randomly) associated with 2/3 of the observations, while a 10^{-10} weight is used for the remaining observations.

The IAs used to evaluate agreement between partitions are the Adjusted Rand Index (Hubert and Arabie 1985) and the Variation of Information (Meila 2007).

4 Data Analysis

Four data sets are used to illustrate the proposed approach: Iris, Wine, Haberman and Diabetes (UCI repository, Asunción and Newman 2007).

Results from the analysis are presented in Tables 2 and 3. In these tables, KM refers to K-Means results and EM to Expectation-Maximization results. The index 0 refers to the original partition, based on the entire dataset (the one being evaluated). Each IA average, standard deviation and coefficient of variation refers to 20 weighted samples' clustering results, providing measures to evaluate stability. The

Table 2 Illustrating the trade-off bias-variability

Results using KM			Results using EM		
(KM_0, weighted samples)	AR	VI	(EM_0, weighted samples)	AR	VI
Average	0.936	0.198	Average	0.819	0.463
Std. deviation	0.086	0.202	Std. deviation	0.120	0.252
Coef. of variation	0.092	1.017	Coef. of variation	0.147	0.545
(KM_0, **Iris**)	0.730	0.760	(EM_0, **Iris**)	0.834	0.556
Average	0.940	0.243	Average	0.906	0.306
Std. deviation	0.033	0.123	Std. deviation	0.077	0.220
Coef. of variation	0.035	0.505	Coef. of variation	0.084	0.720
(KM_0, **Wine**)	0.880	0.437	(EM_0, **Wine**)	0.898	0.390

Table 3 When stability is not a good indicator of agreement with the real partition

Results using KM			Results using EM		
(KM_0, weighted samples)	AR	VI	(EM_0, weighted) samples)	AR	VI
Average	0.950	0.093	Average	0.656	0.700
Std. deviation	0.074	0.120	Std. deviation	0.176	0.236
Coef. of variation	0.078	1.294	Coef. of variation	0.268	0.337
(KM_0, **Haberman**)	0.128	1.297	(EM_0, **Haberman**)	0.141	1.637
Average	0.889	0.246	Average	0.442	0.959
Std. deviation	0.101	0.166	Std. deviation	0.191	0.328
Coef. of variation	0.114	0.674	Coef. of variation	0.431	0.342
(KM_0, **Diabetes**)	0.075	1.604	(EM_0, **Diabetes**)	0.013	1.805

agreement between the clustering solution obtained and the real partition measures the bias (the datasets' names are used to indicate the known (real) classes).

As expected, the analysis results show a clear relationship between the Variation of Information index and the Adjusted Rand index values: the Pearson correlation coefficient values range from -0.974 to -1.

As observed in an earlier work (Cardoso et al. 2008), the IA values yielded by the resampling procedure (measuring variability/stability) tend to overestimate the goodness of the clustering solutions. In addition, there are some situations when stability is clearly not a good indicator of agreement with the real partition: these are the cases illustrated in Table 3.

In an attempt to differentiate these cases, the quality index Calinski and Harabasz (1974) was used for measuring the quality (compactness-separability) of the partitions: the larger the CH value the better the partition. This pseudo-F-statistic has evidenced a good performance in experiments performed by Milligan and Cooper (1985).

Table 4 Calinski and Harabask index values

	Iris	Wine	Haberman	Diabetes
True	**486.321**	**68.252**	**8.36**	**24.299**
KM_0	560.400	70.688	76.342	961.218
KM in weighted samples (average)	555.658	70.499	76.077	954.064
EM_0	524.689	68.370	48.158	179.437
EM in weighted samples (average)	522.365	67.929	36.662	488.330

$$CH\left(\Pi^K\right) = \frac{B\left(\Pi^K\right)/(K-1)}{W\left(\Pi^K\right)/(n-K)}$$

According to the obtained results, the CH index values clearly differentiate data sets in Tables 2 and 3: the real Diabetes and Haberman partitions are not compact and well separated, while the clustering algorithms (K-means in particular) tend to yield solutions that exhibit these properties (see Table 4).

5 Discussion and Perspectives

The proposed cross-validation approach to evaluate clustering solutions has some important advantages:

- It is applicable to all clustering algorithms;
- The sample dimension is not a severe limitation for implementing clustering stability evaluation, since the IA values are based on the entire (weighted) sample, and not in a holdout sample;
- The fact that a weighted sample is used mimics the subsample random drawing, but there is no need to select a classifier to implement cross-validation, which could be less adequate to induce partitions and certainly more time consuming.

Regarding the bias-variability relationship, it should be further investigated. Experiments with additional datasets are necessary to analyze clustering variability/stability as addressed by the proposed approach. The relationship between stability and compactness-separability should also be further evaluated (the present work indicates that the effectiveness of stability indicators may be related with the compactness-separability properties). A complementary approach could be focused on cluster-wise stability (Hennig 2007). In addition, the relationship between stability and the adequacy of the clustering criteria to the data set at hand should be further discussed (recent works in the field, like Ben-David and von Luxburg (2008), should be considered).

Acknowledgements The authors would like to thank the support received from CNPq, FAPESP and UNIDE.

References

Asunción, A., & Newman, D. J. (2007). *UCI machine learning repository*. [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

Ben-David, S., & von Luxburg, U. (2008). Relating clustering stability to properties of cluster boundaries. In R. Servedio & T. Zhang (Eds.), *Proceedings of the 21st Annual Conference on Learning Theory (COLT)* (pp. 379–390). Berlin: Springer.

Breckenridge, J. N. (1989). Replicating cluster analysis: Method, consistency, and validity. *Multivariate Behavioral Research, 24*, 147–161.

Brun, M., Sima, C., Hua, J., Lowey, J., Carrol, B., Suh, E., & Dougherty, E. R. (2007). Model-based evaluation of clustering validation measures. *Pattern recognition, 40*, 807–824.

Calinski, T., & Harabasz, J. (1974). A dendrit method for cluster analysis. *Communications in Statistics, 3*, 1–27.

Cardoso, M. G. M. S. (2007). Clustering and cross-validation. C. L. C. Ferreira, G. Saporta & M. Souto de Miranda (Eds.), *IASC 07 – Statistics for data mining, learning and knowledge extraction*.

Cardoso, M. G. M. S., & Carvalho, A. P. L. F. (2009). Quality indices for (practical) clustering evaluation. *Intelligent Data Analysis, 13*(5), 725–740.

Cardoso, M. G. M. S., de Carvalho, A. P. L. F., & Faceli, K. (2008). Clustering stability and resampling. P. Brito (Ed.), *18th International Conference on Computational Statistics*.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence, 1*(2), 224–227.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistics Society, 39*(1), 1–38. Harvard University and Educational Testing Service.

Dudoit, S., & Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a data set. *Genome Biology, 3*(7), 1–21.

Gordon, A. D. (1999). *Classification*. London: Chapman & Hall/CRC.

Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics and Data Analysis, 52*, 258–271.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*, 193–218.

Lange, T., Roth, V., Braun, M. L., & Buchman, J. M. (2004). Stability based validation of clustering solutions. *Neural Computation, 16*, 1299–1323.

Law, M. H., & Jain, A. K. (2003). *Cluster validity by bootstrapping partitions*. Technical report MSU-CSE-03-5, Department of Computer Science and Engineering. Michigan State University.

Levine, E., & Domany, E. (2001). Resampling method for unsupervised estimation of cluster validity. *Neural Computation, 13*, 2573–2593.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *5th Berkeley Symposium on Mathematical Statistics and Probability*. California: University of California Press.

McIntyre, R. M., & Blashfield, R. K. (1980). A nearest-centroid technique for evaluating the minimum-variance clustering procedure. *Multivariate Behavioral Research, 15*, 225–238.

Meila, M. (2007). Comparing clusterings-an information based distance. *Journal of Multivariate Analysis, 98*(5), 873–895.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50*, 159–179.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66*, 846–850.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, B, 36*(1), 111–147.

Tibshirani, R., Walther, G., Botstein, D., & Brown, P. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics, 14*(3), 511–528.

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89–106). Cambridge: Cambridge University Press.

Cluster Structured Multivariate Probability Distribution with Uniform Marginals

Andrzej Sokolowski and Sabina Denkowska

Abstract Reducing the dimension of classification space we doubt if this transformation is spoiling the original group structure of the mixture generating our data. This problem can be discussed on the basis of analysis of data generated by the specific probability distributions. Denkowska and Sokolowski (1997) proposed two-dimensional probability distribution with probability distributed uniformly on two squares which size is controlled by one parameter. In this paper, the special multivariate probability distribution is proposed. In n-dimensions it has a cluster structure consisting of hypercubes but all marginal distributions are uniform.

1 Introduction

Denkowska and Sokolowski (1997) proposed two-dimensional probability distribution with probability distributed uniformly on two squares which size is controlled by one parameter. In this paper, the special multivariate probability distribution is proposed. In n-dimensions it has a cluster structure consisting of hypercubes. The main interesting feature of this distribution is the fact that its every marginal distribution is uniform. It means that the group structure is completely lost while considering any marginal distribution. In other words, omitting one variable out of n, loses group structure.

2 The 3-Dimensional Case

The density function of the proposed distribution in the three-dimensional case is following:

$$f(x,y,z) = \begin{cases} \frac{1}{a^3+3a(b-a)^2} & \text{if } (x,y,z) \in \mathbf{A} \\ 0 & \text{if } (x,y,z) \notin \mathbf{A} \end{cases}$$

A. Sokolowski (✉)
Department of Statistics, Cracow University of Economics, Poland
e-mail: sokolows@uek.krakow.pl

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_22,

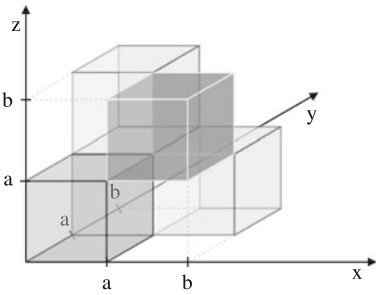


Fig. 1 Domain **A** for which the density function is greater than zero, consists of four prisms (cubes when $b = 2a$)

where:
$\mathbf{A} := (u, u+a)\times(v, v+a)\times(w, w+a)\cup\ (u+a, u+b)\times(v+a, v+b)\times(w, w+a)\cup$
$\cup(u, u+a)\times(v+a, v+b)\times(w+a, w+b)\cup\ (u+a, u+b)\times(v, v+a)\times(w+a, w+b)$
and $(u, v, w) \in R^3, 0 < a < b$.
Set **A** for which the density function is greater than zero, consists of four rectangular prisms (Fig. 1).

2.1 Description of Proposed Probability Distribution

2.1.1 Marginal Distributions

The marginal distributions of proposed three-dimensional probability distribution are uniform in the case when $b = 2a$ and it means that the group structure is lost while considering any marginal distribution.
The density functions of all marginal distributions are listed below:

- The marginal distribution of random vector (X, Y):

$$f_{X,Y}(x, y) = \begin{cases} \frac{a}{a^3+3a(b-a)^2} & \text{if } (x, y) \in (u, u+a) \times (v, v+a) \cup \\ & \cup(u+a, u+b) \times (v+a, v+b) \\ \frac{b-a}{a^3+3a(b-a)^2} & \text{if } (x, y) \in (u, u+a) \times (v+a, v+b) \cup \\ & \cup(u+a, u+b) \times (v, v+a) \\ 0 & \text{otherwise} \end{cases}$$

- The marginal distribution of random vector (Y, Z):

$$f_{Y,Z}(y, z) = \begin{cases} \frac{a}{a^3+3a(b-a)^2} & \text{if } (y, z) \in (v, v+a) \times (w, w+a) \cup \\ & \cup(v+a, v+b) \times (w+a, w+b) \\ \frac{b-a}{a^3+3a(b-a)^2} & \text{if } (y, z) \in (v, v+a) \times (w+a, w+b) \cup \\ & \cup(v+a, v+b) \times (w, w+a) \\ 0 & \text{otherwise} \end{cases}$$

- The marginal distribution of random vector (X, Z):

$$f_{X,Z}(x,z) = \begin{cases} \frac{a}{a^3+3a(b-a)^2} & \text{if } (x,z) \in (u,u+a)\times(w,w+a) \cup \\ & \cup (u+a,u+b)\times(w+a,w+b) \\ \frac{b-a}{a^3+3a(b-a)^2} & \text{if } (x,z) \in (u,u+a)\times(w+a,w+b) \cup \\ & \cup (u+a,u+b)\times(w,w+a) \\ 0 & \text{otherwise} \end{cases}$$

- The marginal distribution of random variable X:

$$f_X(x) = \begin{cases} \frac{a^2+(b-a)^2}{a^3+3a(b-a)^2} & \text{if } x \in (u,u+a) \\ \frac{2a(b-a)}{a^3+3a(b-a)^2} & \text{if } x \in (u+a,u+b) \\ 0 & \text{otherwise} \end{cases}$$

- The marginal distribution of random variable Y:

$$f_Y(y) = \begin{cases} \frac{a^2+(b-a)^2}{a^3+3a(b-a)^2} & \text{if } y \in (v,v+a) \\ \frac{2a(b-a)}{a^3+3a(b-a)^2} & \text{if } y \in (v+a,v+b) \\ 0 & \text{otherwise} \end{cases}$$

- the marginal distribution of random variable Z:

$$f_Z(z) = \begin{cases} \frac{a^2+(b-a)^2}{a^3+3a(b-a)^2} & \text{if } z \in (w,w+a) \\ \frac{2a(b-a)}{a^3+3a(b-a)^2} & \text{if } z \in (w+a,w+b) \\ 0 & \text{otherwise} \end{cases}$$

2.1.2 Raw and Central Moments

Some other characteristics of the proposed distribution are:

- the "$r_1 + r_2 + r_3$"th raw moment ($r_i \in N_0, i = 1,2,3$) equals:

$$m_{r_1 r_2 r_3} = E(X^{r_1} Y^{r_2} Z^{r_3}) = \frac{1}{(r_1+1)(r_2+1)(r_3+1)(a^3+3a(b-a)^2)} \cdot \Big[\Big((w+a)^{r_3+1} - w^{r_3+1}\Big)\Big(\big((v+a)^{r_2+1} - v^{r_2+1}\big)\big((u+a)^{r_1+1} - u^{r_1+1}\big) + \big((v+b)^{r_2+1} - (v+a)^{r_2+1}\big)\big((u+b)^{r_1+1} - (u+a)^{r_1+1}\big)\Big) + \Big((w+b)^{r_3+1} - (w+a)^{r_3+1}\Big)\Big(\big((v+b)^{r_2+1} - (v+a)^{r_2+1}\big)\big((u+a)^{r_1+1} - u^{r_1+1}\big) + \big((v+a)^{r_2+1} - v^{r_2+1}\big)\big((u+b)^{r_1+1} - (u+a)^{r_1+1}\big)\Big)\Big]$$

- the "$r_1 + r_2 + r_3$"th central moment ($r_i \in N_0, i = 1, 2, 3$) equals:

$$M_{r_1 r_2 r_3} = E[(X - m_{100})^{r_1}(Y - m_{010})^{r_2}(Z - m_{001})^{r_3}] =$$
$$\frac{1}{(r_1+1)(r_2+1)(r_3+1)\left(a^3+3a(b-a)^2\right)}\Bigg[\left((w + a - m_{001})^{r_3+1} - (w - m_{001})^{r_3+1}\right)\cdot$$
$$\Bigg(\left((u+a-m_{100})^{r_1+1}-(u-m_{100})^{r_1+1}\right)\left((v+a-m_{010})^{r_2+1}-(v-m_{010})^{r_2+1}\right)+$$
$$\left((u + b - m_{100})^{r_1+1} - (u + a - m_{100})^{r_1+1}\right)\left((v + b - m_{010})^{r_2+1} - (v + a - m_{010})^{r_2+1}\right)\Bigg) + \left((w + b - m_{001})^{r_3+1} - (w + a - m_{001})^{r_3+1}\right)\Bigg(\left((u + b - m_{100})^{r_1+1}-(u+a-m_{100})^{r_1+1}\right)\left((v+a-m_{010})^{r_2+1}-(v-m_{010})^{r_2+1}\right)+\left((u+a-m_{100})^{r_1+1}-(u-m_{100})^{r_1+1}\right)\left((v+b-m_{010})^{r_2+1}-(v+a-m_{010})^{r_2+1}\right)\Bigg)\Bigg]$$

In particular, the expected values and variations of individual random variables are as follows:

$$m_{100} = E(X) =$$
$$= \frac{1}{2\left(a^3+3a(b-a)^2\right)}\left(\left(a^2+(b-a)^2\right)\left((u+a)^2-u^2\right)\right)+2a(b-a)\left((u+b)^2-(u+a)^2\right)\Big)$$
$$m_{010} = E(Y) =$$
$$= \frac{1}{2\left(a^3+3a(b-a)^2\right)}\left(\left(a^2+(b-a)^2\right)\left((v+a)^2-v^2\right)\right)+2a(b-a)\left((v+b)^2-(v+a)^2\right)\Big)$$
$$m_{001} = E(Z) =$$
$$= \frac{1}{2\left(a^3+3a(b-a)^2\right)}\left(\left(a^2+(b-a)^2\right)\left((w+a)^2-w^2\right)\right)+2a(b-a)\left((w+b)^2-(w+a)^2\right)\Big)$$

$$M_{200} = D^2(X) = \frac{\left(a^2 + (b-a)^2\right)\left((u + a - m_{100})^3 - (u - m_{100})^3\right)}{3\left(a^3 + 3a(b-a)^2\right)}+$$
$$+ \frac{2a(b-a)\left((u + b - m_{100})^3 - (u + a - m_{100})^3\right)}{3\left(a^3 + 3a(b-a)^2\right)}$$
$$M_{020} = D^2(Y) = \frac{\left(a^2 + (b-a)^2\right)\left((v + a - m_{010})^3 - (v - m_{010})^3\right)}{3\left(a^3 + 3a(b-a)^2\right)}+$$
$$+ \frac{2a(b-a)\left((v + b - m_{010})^3 - (v + a - m_{010})^3\right)}{3\left(a^3 + 3a(b-a)^2\right)}$$
$$M_{002} = D^2(Z) = \frac{\left(a^2 + (b-a)^2\right)\left((w + a - m_{001})^3 - (w - m_{001})^3\right)}{3\left(a^3 + 3a(b-a)^2\right)}+$$
$$+ \frac{2a(b-a)\left((w + b - m_{001})^3 - (w + a - m_{001})^3\right)}{3\left(a^3 + 3a(b-a)^2\right)}$$

2.1.3 The Characteristic Function

To obtain a full formal description of the proposed distribution, the characteristic function is derived:

$$\varphi(t_1,t_2,t_3)=E(e^{i(t_1X+t_2Y+t_3Z)})=\frac{1}{-it_1t_2t_3(a^3+3a(b-a)^2)}\big[(e^{it_1(u+a)}-e^{it_1u})\cdot$$

$$(e^{it_2(v+a)}-e^{it_2v})(e^{it_3(w+a)}-e^{it_3w})+(e^{it_1(u+b)}-e^{it_1(u+a)})(e^{it_2(v+b)}-e^{it_2(v+a)})\cdot$$

$$(e^{it_3(w+a)}-e^{it_3w})+(e^{it_1(u+a)}-e^{it_1u})(e^{it_2(v+b)}-e^{it_2(v+a)})(e^{it_3(w+b)}-e^{it_3(w+a)})+$$
$$(e^{it_1(u+b)}-e^{it_1(u+a)})(e^{it_2(v+a)}-e^{it_2v})(e^{it_3(w+b)}-e^{it_3(w+a)})\big]$$

3 Cluster Structured 3-Dimensional Distribution with Uniform Marginals

To simplify, let's assume that $(u,v,w)=(0,0,0)$.
The most interesting case is when $b=2a$. Then each marginal distribution is uniform. Then probability density function for random variable (X, Y, Z) is the following:

$$f(x,y,z)=\begin{cases}\frac{1}{4a^3} & \text{if}(x,y,z)\in\mathbf{A}\\ 0 & \text{if } (x,y,z)\notin\mathbf{A}\end{cases}$$

where
$A=(0,a)\times(0,a)\times(0,a)\cup(a,2a)\times(a,2a)\times(0,a)\cup(0,a)\times(a,2a)\times(a,2a)\cup$
$(a,2a)\times(0,a)\times(a,2a)$.

In this case the selected raw and central moments are:
$E(X)=E(Y)=E(Z)=a$
$D^2(X)=D^2(Y)=D^2(Z)=\frac{a^2}{3}$.
The marginal distribution of random vector (X, Y) is uniform inside the square with the side (0, 2a) without line segments connecting the centers of the opposite sides:

$$f_{X,Y}(x,y)=\begin{cases}\frac{1}{4a^2} & \text{if } (x,y)\in(0,a)\times(0,a)\cup(a,2a)\times(a,2a)\cup\\ & \cup(0,a)\times(a,2a)\cup(a,2a)\times(0,a)\\ 0 & \text{otherwise}\end{cases}$$

The density functions for the remaining cases of the two-dimensional random vectors are analogical:

$$f_{Y,Z}(y,z)=\begin{cases}\frac{1}{4a^2} & \text{if } (y,z)\in(0,a)\times(0,a)\cup(a,2a)\times(a,2a)\cup\\ & \cup(0,a)\times(a,2a)\cup(a,2a)\times(0,a)\\ 0 & \text{otherwise}\end{cases}$$

$$f_{X,Z}(x,z)=\begin{cases}\frac{1}{4a^2} & \text{if } (x,z)\in(0,a)\times(0,a)\cup(a,2a)\times(a,2a)\cup\\ & \cup(0,a)\times(a,2a)\cup(a,2a)\times(0,a)\\ 0 & \text{otherwise}\end{cases}$$

The marginal distribution of every individual random variable is uniform on the segment $(0, 2a)$ without its center:

$$f_X(x) = \begin{cases} \frac{1}{2a} & \text{if } x \in (0,a) \cup (a,2a) \\ 0 & \text{otherwise} \end{cases}$$
$$f_Y(y) = \begin{cases} \frac{1}{2a} & \text{if } y \in (0,a) \cup (a,2a) \\ 0 & \text{otherwise} \end{cases}$$
$$f_Z(z) = \begin{cases} \frac{1}{2a} & \text{if } z \in (0,a) \cup (a,2a) \\ 0 & \text{otherwise} \end{cases}$$

Both in the case of two-dimensional and one-dimensional probability distributions, the exclusion of, respectively, the sum of segments or the point of the domain for which the density function is greater than zero, does not influence the properties of uniform distribution.

To simplify, let's consider the case when $a = \frac{1}{2}, b = 1$. Then the probability function is as follows:

$$f(x,y,z) = \begin{cases} 2 & \text{if } (x,y,z) \in (0,\frac{1}{2}) \times (0,\frac{1}{2}) \times (0,\frac{1}{2}) \cup (\frac{1}{2},1) \times (\frac{1}{2},1) \times (0,\frac{1}{2}) \cup \\ & \cup (0,\frac{1}{2}) \times (\frac{1}{2},1) \times (\frac{1}{2},1) \cup (\frac{1}{2},1) \times (0,\frac{1}{2}) \times (\frac{1}{2},1) \\ 0 & \text{otherwise} \end{cases}$$

In this case expected values and variations are:
$E(X) = E(Y) = E(Z) = \frac{1}{2}$
$D^2(X) = D^2(Y) = D^2(Z) = \frac{1}{12}$
and all marginal distributions are uniform:

$$f_{X,Y}(x,y) = \begin{cases} 1 & \text{if } (x,y) \in (0,\frac{1}{2}) \times (0,\frac{1}{2}) \cup (\frac{1}{2},1) \times (\frac{1}{2},1) \cup \\ & \cup (0,\frac{1}{2}) \times (\frac{1}{2},1) \cup (\frac{1}{2},1) \times (0,\frac{1}{2}) \\ 0 & \text{otherwise} \end{cases}$$
$$f_{X,Z}(x,z) = \begin{cases} 1 & \text{if } (x,z) \in (0,\frac{1}{2}) \times (0,\frac{1}{2}) \cup (\frac{1}{2},1) \times (\frac{1}{2},1) \cup \\ & \cup (0,\frac{1}{2}) \times (\frac{1}{2},1) \cup (\frac{1}{2},1) \times (0,\frac{1}{2}) \\ 0 & \text{otherwise} \end{cases}$$
$$f_{Y,Z}(y,z) = \begin{cases} 1 & \text{if } (y,z) \in (0,\frac{1}{2}) \times (0,\frac{1}{2}) \cup (\frac{1}{2},1) \times (\frac{1}{2},1) \cup \\ & \cup (0,\frac{1}{2}) \times (\frac{1}{2},1) \cup (\frac{1}{2},1) \times (0,\frac{1}{2}) \\ 0 & \text{otherwise} \end{cases}$$

$$f_X(x) = \begin{cases} 1 & \text{if } x \in (0,\frac{1}{2}) \cup (\frac{1}{2},1) \\ 0 & \text{otherwise} \end{cases}$$
$$f_Y(y) = \begin{cases} 1 & \text{if } y \in (0,\frac{1}{2}) \cup (\frac{1}{2},1) \\ 0 & \text{otherwise} \end{cases}$$
$$f_Z(z) = \begin{cases} 1 & \text{if } z \in (0,\frac{1}{2}) \cup (\frac{1}{2},1) \\ 0 & \text{otherwise} \end{cases}$$

4 The *n*-Dimensional Case

Three-dimensional (X, Y, Z) distribution can be generalized onto n-dimensional case. The main problem lies in finding such configuration of hypercubes for which marginal distributions are uniform.

To simplify, let's assume that $b = 2a$ (in that case all marginals are uniform) and let's use the notation which shortens the description of hypercubes for which the density function is non-zero. So, in case when $n = 2$, the notation "**01**" means that the density function is greater than zero for $(x, y) \in (u, u+a) \times (v+a, v+b)$. The first digit informs us about the first coordinate, and "**0**" means that $x \in (u, u+a)$, and "**1**" that $x \in (u+a, u+b)$; the second digit informs about the second coordinate, and, analogically, "**0**" means that $y \in (v, v+a)$, and "**1**" means that $y \in (v+a, v+b)$. In the two-dimensional case described in [1], using proposed notation, squares with non-zero density function would be labeled: "**00**", "**11**". In the case of probability distribution of random vector (X,Y,Z) proposed in this paper, the density function equals $\frac{1}{a^3+3a(b-a)^2}$ for four cubes: "**000**", "**110**", "**101**" and "**011**". Obviously a proposed choice of cubes isn't the only solution, but it is convenient to extending to n-dimensional case.

The considered n-dimensional hypercubes (for $n = 2, 3, 4, 5$) are listed in Table 1, and those with non-zero density function are reported in bold. For two and three dimensional space the choice of hypercubes is easy, but in four-dimensional space it is not so obvious. Let's have a look how the cubes have been described and chosen in the low-dimensional spaces. To move from the two-dimensional distribution to the three-dimensional one, the **0/1** sequence from the first column (for $n = 2$) had to be expanded by the third coordinate first with **0**'s and then with **1**'s. In that way, the description of squares has been expanded into the description of cubes. For the first four cubes $z \in (w, w+a)$, and for the rest of them $z \in (w+a, w+b)$. The choice of cubes with non-zero density function is given in Table 1. When the last coordinate equals **0**, we chose the same sequence of **0/1** as for the lower dimension, and finally, every second cube. Then the rest of the cubes are chosen among those with **1** in the last coordinate.

Having $n-1$ cubes in $(n-1)$-dimensional space, which form set **A** with non-zero density function, we can construct the distribution of n-dimensional random variable. Simplifying the problem by assuming $b = 2a$ (in that case all marginals are uniform) we come to the following density function:

$$f(x_1, \ldots, x_n) = \begin{cases} \frac{1}{2^{n-1}a^n} & \text{if } (x_1, \ldots, x_n) \in \mathbf{A} \\ 0 & \text{if } (x_1, \ldots, x_n) \notin \mathbf{A} \end{cases}$$

A random number generator has been written in STATISTICA Visual Basic (SVB) for the proposed distribution. Generated data sets have been analysed by various multidimensional methods provided in STATISTICA. The results show the expected uniformity of marginal distributions and lack of group structure while reducing the dimensionality of classification space.

Table 1 The n-dimensional hypercubes (for $n = 2, 3, 4, 5$). Domain **A** for which the density function is greater than zero consists of cubes reported in bold

$n = 2$	$n = 3$	$n = 4$	$n = 5$
0 0	**0 0 0**	**0 0 0 0**	**0 0 0 0 0**
1 0	1 0 0	1 0 0 0	1 0 0 0 0
1 1	**1 1 0**	**1 1 0 0**	**1 1 0 0 0**
0 1	0 1 0	0 1 0 0	0 1 0 0 0
	0 0 1	**0 0 1 0**	**0 0 1 0 0**
	1 0 1	1 0 1 0	1 0 1 0 0
	1 1 1	**1 1 1 0**	**1 1 1 0 0**
	0 1 1	0 1 1 0	0 1 1 0 0
		0 0 0 1	**0 0 0 1 0**
		1 0 0 1	1 0 0 1 0
		1 1 0 1	**1 1 0 1 0**
		0 1 0 1	0 1 0 1 0
		0 0 1 1	**0 0 1 1 0**
		1 0 1 1	1 0 1 1 0
		1 1 1 1	**1 1 1 1 0**
		0 1 1 1	0 1 1 1 0
			0 0 0 0 1
			1 0 0 0 1
			1 1 0 0 1
			0 1 0 0 1
			0 0 1 0 1
			1 0 1 0 1
			1 1 1 0 1
			0 1 1 0 1
			0 0 0 1 1
			1 0 0 1 1
			1 1 0 1 1
			0 1 0 1 1
			0 0 1 1 1
			1 0 1 1 1
			1 1 1 1 1
			0 1 1 1 1

Reference

Denkowska, S., & Sokolowski A. (1997). O pewnym rozkladzie prawdopodobienstwa dwuwymiarowej zmiennej losowej (On a certain probability distribution of a two-dimensional random variable). Cracow University of Economics, *Proceedings of 35th conference of statistics, econometrics and mathematics departments of universities of economics* (pp. 59–67). Cracow.

Analysis of Diversity-Accuracy Relations in Cluster Ensemble

Dorota Rozmus

Abstract Ensemble approaches based on aggregated models have been successfully applied in the context of supervised learning in order to increase the accuracy and stability of classification. Recently, analogous techniques for cluster analysis have been introduced. Research has proved that, by combining a collection of different clusterings, an improved solution can be obtained.

Diversity within an ensemble is very important for its success. An ensemble of identical classifiers or clusterers will not be better than the individual ensemble members. However, finding a sensible quantitative measure of diversity in classifier ensembles turned out to be very difficult (Kuncheva 2003; Kuncheva and Whitaker 2003). Diversity in cluster ensembles is considered here. The aim of the research is to look into the relationship between diversity and the accuracy of the cluster ensemble.

1 Introduction

Ensemble techniques based on aggregated models have been successfully applied in supervised learning in order to improve the accuracy and stability of classification algorithms (Breiman 1996). The idea of aggregation can be formulated as follows: instead one model for prediction use many different models and then combine many predicted values with some aggregation operator. In the case of classification the most often used operator is majority voting, i.e. an observation is assigned to the most often chosen class. Among the most popular methods there are eg. bagging based on bootstrap samples (Breiman 1996) and boosting based on increasing weights to the wrongly classified examples (Freund 1990).

In the last few years, the ensemble approach for cluster analysis has been introduced. It is believed that it allows to increase the classification accuracy and

D. Rozmus
Department of Statistics, Katowice University of Economics, Bogucicka 14, 40-226 Katowice
e-mail: drozmus@ae.katowice.pl

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_23,

robustness of the clustering solutions. The main aim of aggregation is to combine results of several different clusterings in order to get a final clustering with better quality. Recently several studies on combination methods of clustering solutions have established a new area in the conventional taxonomy. There are several possible ways of applying the idea of ensemble approach in the context of unsupervised learning (1) aggregation of results of different clustering algorithms; (2) receiving different partitions by resampling the data, such as in bootstrapping techniques, eg. bagging; (3) applying different subsets of features (disjoint or overlapping); (4) using a given algorithm many times with different values of parameters or initializations.

2 Diversity Measures

Intuitively, an ensemble works the best when its clustering solutions are of good quality and at the same time differ from one another significantly. Diversity within an ensemble is very important for its success. If all the clusterers or classifiers of an ensemble will be the same then the aggregated solution will not outperform the individual ensemble members. However, finding a good quantitative measure of diversity in classifier ensembles has appeared very difficult (Kuncheva 2005, 2003; Kuncheva and Whitaker 2003). In last few years, diversity of cluster ensembles has also been studied. Fern and Brodley (2003) have found that ensemble members with higher diversity offer larger improvement than less diverse ensembles. In the literature many different diversity measures for cluster ensembles have been suggested. In this work the concept of measure introduced by Hadjitodorov et al. (2006) was used. The Authors presented five measures based on accuracy index of the clustering algorithm. In this proposition the measure of diversity and the match index for the ensemble accuracy was based on the Adjusted Rand Index. But from experiments, carried out by mentioned Authors, the relationship between diversity measures based on this index and the accuracy of the ensemble was not quite clear. The aim of this work is to check if applying of other popular accuracy indexes as a base for proposed diversity measures will give any improvement, i.e. will reveal clear relationship between diversity measures and the quality of ensemble so that one can pick from a set of ensembles the one that is most likely to be good. In this work Rand Index, Jaccard Index and Fowlkes and Mallows Index were used. Rand Index which measures a similarity between the final (aggregated) partition C_{agr} and the true labels C^T as a measure of correctness of the final partition was used.

Generally there are two ways for measuring the ensemble diversity: pairwise and non-pairwise approach. In the first approach the ensemble diversity measure proposed by Hadjitodorov, Kuncheva and Todorova is given by formula:

$$D_p = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} (1 - acc(C_i, C_j)) \tag{1}$$

where as a measure of accuracy (*acc*) the Authors used Adjusted Rand Index, and in this research Rand Index, Jaccard Index and Fowlkes and Mallows Index was used. It should be noted that *acc* gives similarity between partitions, therefore $1 - acc(C_i, C_j)$ would be the measure of pairwise diversity.

For the non-pairwise approach, after the ensemble decision is specified, each clusterer has assigned a diversity value measuring its difference from the ensemble output. So in order to obtain an overall measure of diversity one can simply take the average of the M individual diversities:

$$D_{np1} = \frac{1}{M}\sum_{i=1}^{M}(1 - acc(C_i, C_{agr})) \tag{2}$$

In the previous studies (Kuncheva and Hadjitodorov 2004) the mentioned Authors have found that ensembles with a larger spread of individual diversities are generally better than ensembles with a smaller spread. Therefore, they constructed the second non-pairwise diversity measure as the standard deviation of the individual diversities:

$$D_{np2} = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}[(1 - acc(C_i, C_{agr})) - D_{np1}]^2} \tag{3}$$

The Authors discovered also that the spread alone was not strongly related to the ensemble accuracy either. Therefore they proposed a third non-pairwise diversity measure that was based on the following intuition. Because it is believed that the ensemble decision is close to the true labeling of the data, the accuracy of the individual clusterers can be estimated based on how close they are to the ensemble decision. Thus larger values of $1 - D_{np1}$ should be preferred. On the other hand, variability within the ensemble can be estimated by the spread of the individual diversities. Large variability will be indicated by larger values of D_{np2}. The simplest compromise measure can be constructed as:

$$D_{np3} = \frac{1}{2}(1 - D_{np1} + D_{np2}) \tag{4}$$

Another compromise measure can be constructed as the coefficient of variation:

$$D_{np4} = \frac{D_{np3}}{D_{np1}} \tag{5}$$

3 Numerical Experiments and Results

In the research artificial data sets were used that are usually applied in comparative studies in taxonomy. Their short characteristics are shown in Table 1. The inputs of the *Cuboids* problem are uniformly distributed on a three-dimensional space within

Table 1 Characteristics of data sets

Data set	# instances	# features	# classes
Cuboids	500	3	4
Smiley	500	2	4
Spirals	500	2	2
Ringnorm	500	2	2
Threenorm	500	2	2

three cuboids and a small cube in the middle of them. The *Smiley* consists of two Gaussian eyes, a trapezoid nose and a parabula mouth (with vertical Gaussian noise). The inputs of the *Spirals* data set are points on two entangled spirals. The inputs of the *Ringnorm* problem are points from two Gaussian distributions. Class 1 is multivariate normal with mean 0 and covariance four times the identity matrix. Class 2 has unit covariance and mean $(a, a, \ldots, a)$, where $a = d^{-0.5}$ and d is dimension of the problem. The inputs of the *Threenorm* data set are points from two Gaussian distributions with unit covariance matrix. Class 1 is drawn with equal probability from a unit multivariate normal with mean $(a, a, \ldots, a)$ and from a unit multivariate normal with mean $(-a, -a, \ldots, -a)$. Class 2 is drawn from a multivariate normal with mean at $(a, -a, a, \ldots, -a), a = 2/d^{-0.5}$. The first three sets have clearly separated classes and the second two are sets with overlapping classes.

A cluster ensemble of length $B = 50$ was built by running the k-means algorithm on bootstrap samples of the data. The final partition was obtained by the *optimization approach* which formalizes the natural idea of describing consensus clusterings as the ones which "optimally represent the ensemble" by providing a criterion to be optimized over a suitable set C of possible consensus clusterings. If *dist* is an Euclidean dissimilarity measure and $(c_1; \ldots; c_B)$ are the elements of the ensemble, the problem is solved by means of *least squares* consensus clusterings (generalized means; Hornik 2005):

$$\sum_{b=1}^{B} dist(c, c_b)^2 \Rightarrow min_{c \in C} \tag{6}$$

Generally, results reveal that very clear and strict relation between the diversity and the accuracy measures there is only in the case of one data set, i.e. for *Threenorm*.

Looking at the diagrams (Fig. 1) where on x-axis there is the accuracy of the ensemble, and on y-axis there is diversity measure, in the case of the pairwise measure it can be observed that the relationship is almost linear and negative directed. For the first non-pairwise measure there can observed very similar behaviour – negative and linear correlation, but it seems to be more strict than for the pairwise measure. The second, third and fourth non-pairwise measures reveal the same patter – linear and positive correlation with greater intensity of the strength than in the case of the pairwise and the first non-pairwise measure.

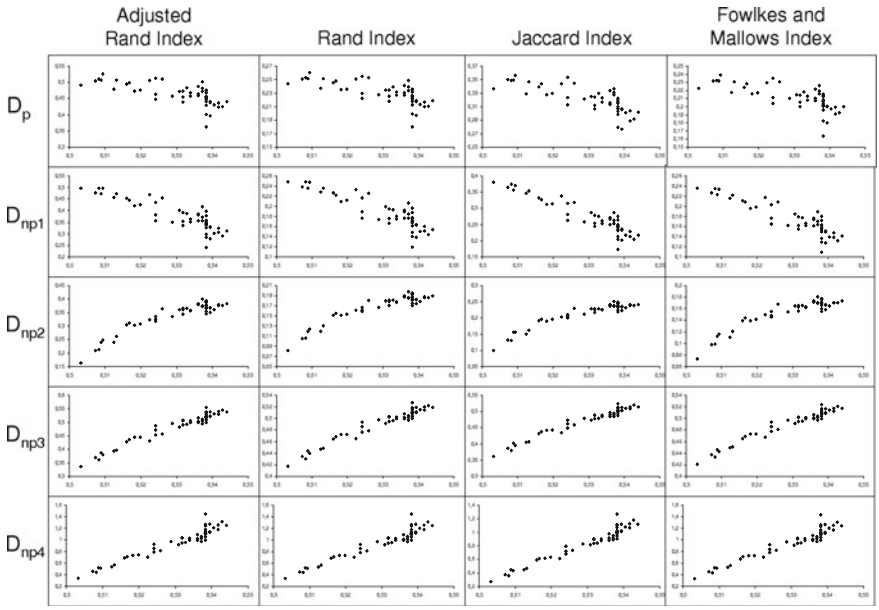


Fig. 1 Accuracy-diversity relationship for the *Threenorm* data set

Table 2 Pearson's linear coefficient of correlation

Diversity measure	Adj. rand index	Rand index	Jaccard index	Fowlkes and mallows index
D_p	−0.679	−0.682	−0.717	−0.693
D_{np1}	−0.869	−0.874	−0.924	−0.891
D_{np2}	0.933	0.931	0.938	0.923
D_{np3}	0.973	0.973	0.978	0.973
D_{np4}	0.957	0.957	0.967	0.958

Because the relationship seems to be linear it is also possible to count Pearson's linear coefficient of correlation in order to check which of the four indexes gave the strongest relationship between accuracy and diversity. From the results presented in the Table 2 it appears that the best is Jaccard Index, especially with the third non-pairwise measure.

Moving towards the next data sets, at the beginning it should be noted that they did not give so strict and clear relationship but in few cases follows some general pattern: lower diversity for higher accuracy of an ensembles in the case of the pairwise and first non-pairwise measure, and for the rest of non-pairwise measures the relation was inverse: higher diversity went together with higher accuracy.

In the case of *Cuboids* data set (Fig. 2) this general lower diversity-higher accuracy pattern can be noticed especially for the first non-pairwise measure. In the case of the rest of non-pairwise measures the third and fourth measure with Jaccard and

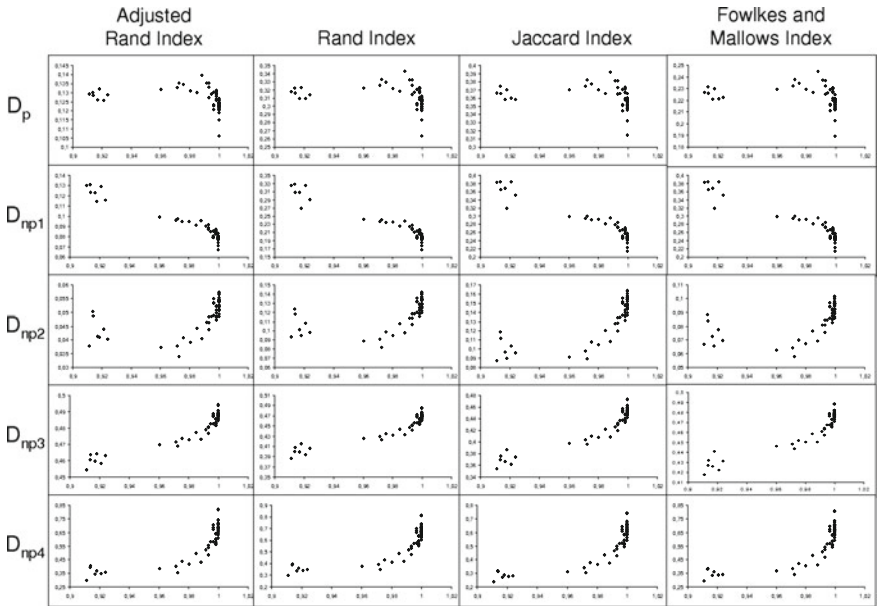


Fig. 2 Accuracy-diversity relationship for the *Cuboids* data set

Fowlkes and Mallows Index especially confirm the general higher diversity-higher accuracy pattern.

For the *Ringnorm* data set (Fig. 3) in the case of non-pairwise and first pairwise measure it is rather difficult to find a confirmation of the general lower diversity – higher accuracy pattern, whereas the rest of non-pairwise measures very slightly confirm the higher diversity – higher accuracy pattern.

For the *Smiley* data set (Fig. 4) again in the case of the non-pairwise and first non-pairwise measure it is difficult to find any clear relationship. For the rest of non-pairwise measures one can conclude that although high diversity sometimes goes together with relatively low accuracy, but high diversity is essential for the high accuracy of the ensemble.

In the case of Spirals data set (Fig. 5) it is rather difficult to find any clear relationship for both pairwise and non-pairwise measures.

4 Summary

To sum up, it should be noted that it is generally believed that diverse cluster ensembles are better than non-diverse ensembles but on the basis of results from discrimination field it is also accepted that the relationship between diversity and accuracy is not clear and straightforward (Kuncheva and Whitaker 2003). Since

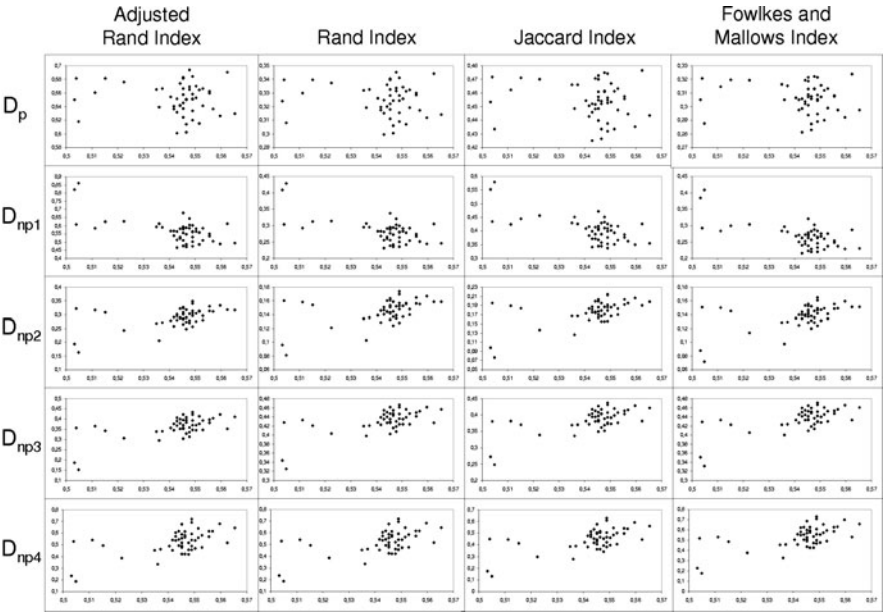


Fig. 3 Accuracy-diversity relationship for the *Ringnorm* data set

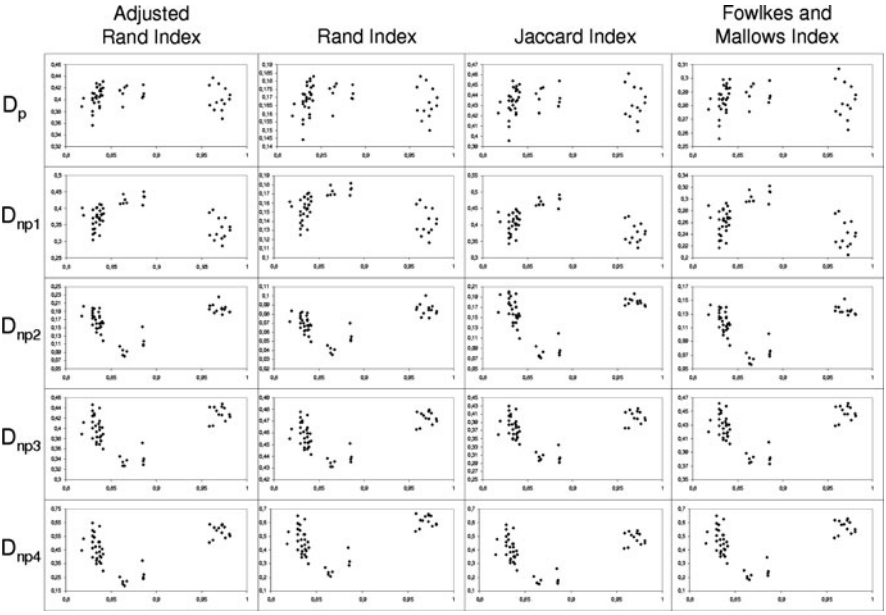


Fig. 4 Accuracy-diversity relationship for the *Smiley* data set

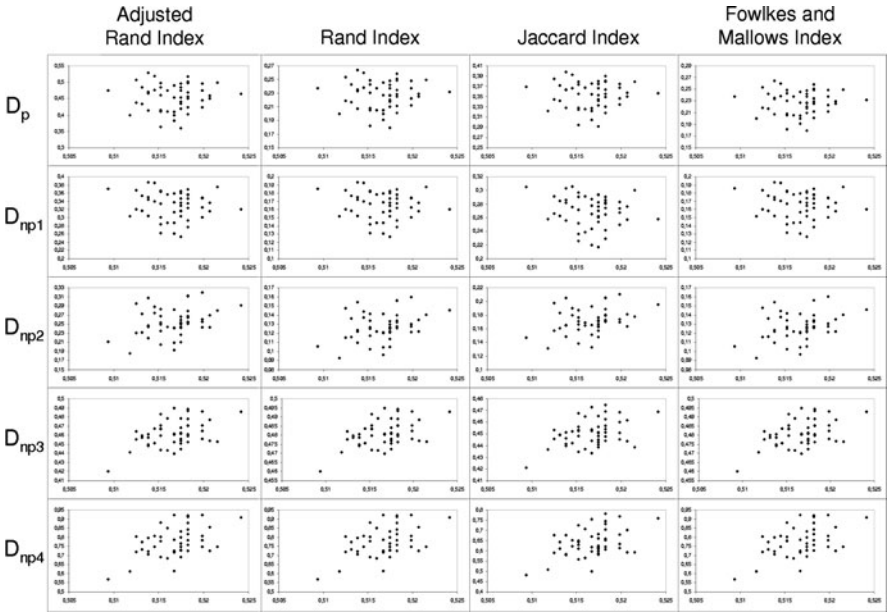


Fig. 5 Accuracy-diversity relationship for the *Spirals* data set

ensemble diversity is a loosely defined concept, there are many different ways to specify and measure it. Here were used five measures proposed in the literature but with different base accuracy measures for estimating diversity in cluster ensembles. From the experiments carried out it is rather difficult to find a strict and clear relationship between ensemble accuracy and the used measures of diversity, but in some cases it can be observed that for the non-pairwise and first non-pairwise measure lower diversity went together with higher accuracy whereas for the rest of non-pairwise measures higher diversity went together with higher accuracy. It is also rather difficult to point out which of those four used indexes gave the strongest relationship between diversity and accuracy. The only exception was the *Threenorm* data set where the relationship was linear and the best index was Jaccard Index.

References

Breiman, L. (1996). Bagging predictors. *Machine Learning, 26*(2), 123–140.

Fern, X. Z., & Brodley, C. E. (2003). Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proceedings of the 20th International Conference on Machine Learning* (pp. 186–193). Washington, DC: ICML.

Freund, Y. (1990). Boosting a weak learning algorithm by majority. In *Proceedings of the 3rd Annual Workshop on Computational Learning Theory* (pp. 202–216).

Hadjitodorov, S. T., Kuncheva, L. I., & Todorova, L. P. (2006). Moderate diversity for better cluster ensembles. *Information Fusion, 7*, 264–275.

Hornik, K. (2005). A clue for cluster ensembles. *Journal of Statistical Software, 14*, 65–72.
Kuncheva, L. I. (2003). That elusive diversity in classifier ensembles. *Lecture Notes in Computer Science* (Vol. 2652, pp. 1126–1138). Mallorca, Spain: Springer-Verlag.
Kuncheva, L. I. (2005). Diversity in multiple classifier systems (Editorial). *Information Fusion, 6*(1), 3–4.
Kuncheva, L., & Hadjitodorov, S. T. (2004). Using diversity in cluster ensembles. *Proceedings of IEEE International Confference on Systems, Man and Cybernetics*.
Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles. *Machine Learning, 51*, 181–207.

Linear Discriminant Analysis with more Variables than Observations: A not so Naive Approach

A. Pedro Duarte Silva

Abstract A new linear discrimination rule, designed for two-group problems with many correlated variables, is proposed. This proposal tries to incorporate the most important patterns revealed by the empirical correlations while approximating the optimal Bayes rule as the number of variables grows without limit. In order to achieve this goal the new rule relies on covariance matrix estimates derived from Gaussian factor models with small intrinsic dimensionality.

Asymptotic results show that, when the model assumed for the covariance matrix estimate is a reasonable approximation to the true data generating process, the expected error rate of the new rule converges to an error close to that of the optimal Bayes rule, even in several cases where the number of variables grows faster than the number of observations.

Simulation results suggest that the new rule clearly outperforms both Fisher's and Naive linear discriminant rules in the data conditions it was designed for.

1 Introduction

The classical theory of Linear Discriminant Analysis (see McLachlan 1992) assumes the existence of a training data set with more observations than variables leading to a non-singular empirical covariance matrix. However, nowadays many applications work with data bases where a large number of variables is measured on a smaller set of observations. Practical experience has shown (Dudoit et al. 2002) that, for problems of this type, natural extensions of Fisher's linear discriminant rule have a disappointing performance. On the other hand, in the same problems the Naive discriminant rule that ignores all variable correlations can be quite effective.

In a seminal paper, Bickel and Levina (2004) have shown that these surprising results have a deep theoretical justification. Using an asymptotic analysis which

A.P.D. Silva
Faculdade de Economia e Gestão & CEGE, Universidade Católica Portuguesa at Porto,
Rua Diogo Botelho, 1327, 4169-005 Porto, Portugal
e-mail: psilva@porto.ucp.pt

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_24,

allows the number of variables to grow faster than the number of observations, these authors have demonstrated that the expected error of the Naive rule can approach a constant close to the expected error of the optimal Bayes rule, while generalized versions of Fisher's rule are asymptotically no better than simple random guessing ignoring the data.

Here, it will be shown that Linear Discriminant Rules based on covariance estimates derived from low-dimensional factor models, can successfully incorporate some of the information available on the empirical correlations and, under conditions similar to those considered in Bickel and Levina (2004), can either achieve or come close to asymptotical optimality, for some problems where both Fisher's and Naive Bayes rules perform poorly.

The reminder of this paper is as follows. In Sect. 2 the new proposal is presented and Sect. 3 discusses its asymptotic properties. Section 4 describes preliminary simulation results and Sect. 5 concludes the paper.

2 A Not so Naive Linear Discriminant Rule

Consider the two-group homoscedastic Gaussian model where entities are represented by binary pairs $(X, Y); X \in \Re^p; Y \in \{0, 1\}$ and the distribution of X conditioned on Y is the multivariate normal $N_p(\mu_{(Y)}, \Sigma)$. The classical discriminant problem deals with the development of rules capable of predicting unknown Y values (class labels) given X observations. When the parameters $\mu_{(0)}, \mu_{(1)}, \Sigma$ are known and the a-priori probabilities $\pi_0 = P(Y = 0)$, $\pi_1 = P(Y = 1)$ are equal it is well known (McLachlan 1992) that the minimal error classification rule is the population Bayes rule

$$Y = \delta_B(X) = \mathbf{1}(\Delta^T \Sigma^{-1} \gamma > 0) \tag{1}$$

where $\Delta = \mu_{(1)} - \mu_{(0)}$ and $\gamma = X - \frac{1}{2}(\mu_{(0)} + \mu_{(1)})$.

In practice, Δ, γ and Σ^{-1} are usually unknown and need to be estimated from a training sample of $n = n_0 + n_1$ observations $((X_i, Y_i); i = 1, \ldots, n)$ with known class labels. When $n_0 \gg p$ and $n_1 \gg p$ common estimators are

$$\hat{\Delta} = \bar{X}_1 - \bar{X}_0 = \frac{1}{n_1} \sum_{Y_i = 1} X_i - \frac{1}{n_0} \sum_{Y_i = 0} X_i \tag{2}$$

$$\hat{\gamma} = X - \frac{1}{2}\left[\bar{X}_0 + \bar{X}_1\right] \tag{3}$$

$$\hat{\Sigma}_F = \frac{1}{n-2}\left[\sum_{Y_i = 0} (X_i - \bar{X}_0)(X_i - \bar{X}_0)^T + \sum_{Y_i = 1} (X_i - \bar{X}_1)(X_i - \bar{X}_1)^T\right] \tag{4}$$

which, for non-singular $\hat{\Sigma}_F$, leads to the Fisher's rule

$$Y = \delta_F(X) = \mathbf{1}(\hat{\Delta}^T \hat{\Sigma}_F^{-1} \hat{\gamma} > 0) \tag{5}$$

Here, we will be most concerned with problems where p is close to, or higher than n. In the latter case $\hat{\Sigma}_F$ is singular and rule (5) can not be applied directly. However, a modified Fisher's rule can be defined by replacing $\hat{\Sigma}_F^{-1}$ by $\hat{\Sigma}_F$ Moore-Penrose generalized inverse.

Alternatively, when Σ is estimated by the diagonal matrix of training sample variances, $\hat{\Sigma}_I = \text{diag}(\hat{\Sigma}_F)$, one gets the Naive Bayes rule

$$Y = \delta_I(X) = \mathbf{1}(\hat{\Delta}^T \ \hat{\Sigma}_I^{-1}\hat{\gamma} > 0) \tag{6}$$

Several studies (see e.g. Dudoit et al. 2002) have shown that rule (6) is surprisingly effective even in problems where many variables are clearly correlated, and when $p \gg n$ often outperforms more sophisticated approaches that do not rely on Gaussian assumptions.

Recently, Bickel and Levina (2004) have demonstrated that under general conditions where $p \to \infty$, $(\ln p)/n \to 0$ and $n/p \to d < \infty$, the worst-possible expected error rate of the modified Fisher's rule converges to $1/2$, but a variant of δ_I that replaces $\hat{\Delta}$ and $\hat{\gamma}$ by consistent estimators of Δ and γ can have a much smaller asymptotic error rate. Furthermore, when the ratio between the largest and the smallest eigenvalues of the population correlation matrix can be bounded by some "moderate" constant, the asymptotic performance of the Naive rule is close to that of the optimal Bayes.

Building on these results, here we will propose an alternative linear rule with good asymptotic performance for some common data conditions where ratios of correlation eigenvalues can be large. In particular, we will assume that true population covariance can be reasonably approximated by a covariance matrix derived from the following q-dimensional ($q \ll p$) factor model

$$X = \mu_{(Y)} + \beta \ F + \Omega \ \epsilon; \beta \in \Re^{p*q}; \Omega = diag(\omega_1, \ldots, \omega_p); \omega_j > k_0 \in \Re_+ \tag{7}$$

where F and ϵ are respectively q-dimensional and p-dimensional random vectors following $N_q(0, I_q)$ and $N_p(0, I_p)$ distributions. When model (7) holds the X covariance matrix, given by $\Sigma = \beta\beta^T + \Omega^2$, is non-singular with inverse equal to

$$\Sigma^{-1} = \Omega^{-2} - \Omega^{-2}\beta[I_q + \beta^T\Omega^{-2}\beta]^{-1}\beta^T\Omega^{-2} \tag{8}$$

This suggests the rule

$$Y = \delta_{Fct_q}(X) = \mathbf{1}(\hat{\Delta}^T \ \hat{\Sigma}_{Fct_q}^{-1}\hat{\gamma} > 0) \tag{9}$$

where $\hat{\Sigma}_{Fct_q}$ is given by

$$\hat{\Sigma}_{Fct_q} = \hat{\beta} \ \hat{\beta}^T + \hat{\Omega}^2; (\hat{\beta}, \hat{\Omega}) = arg\ min||\hat{\Sigma}_{Fct_q} - \hat{\Sigma_F}||^2 \tag{10}$$

and $||\cdot||$ denotes the Frobenius matrix norm, $||A||^2 = tr(A^T A)$.

3 Asymptotic Properties

In this section we will discuss the min-max asymptotic performance of rule δ_{Fct_q}. In particular, we will be concerned with the conditions for convergence, and the limit of the min-max expected misclassification error

$$\overline{W}_{\Gamma_{Fctq}}(\delta_{Fct_q}) = max_{\Gamma_{Fctq}}\left[P_\theta(\delta_{Fct_q}(X) = 1|Y = 0)\right] \tag{11}$$

where

$$\theta = (\Delta, \gamma, \Sigma) \in \Gamma_{Fct_q}(k_0, k_1, k_2, B, c)$$

and

$$\Gamma_{Fct_q}(k_0, k_1, k_2, q, B, c) = \left\{ \begin{array}{c} (\Delta, \gamma, \Sigma): \\ \Delta^T \Sigma^{-1} \Delta \geq c^2 \\ k_1 \leq \lambda_{min}(\Sigma) \leq \lambda_{max}(\Sigma) \leq k_2 \\ \Delta, \gamma \in B \\ \forall\, j = 1, \ldots, p; \quad a = 1, \ldots, q \\ \sum_{j',l'} \left| \frac{\partial \beta(j,a)}{\partial \Sigma(j',l')} \right| \rightarrow e < \infty \\ \sum_{j',l'} \left| \frac{\partial \omega_j}{\partial \Sigma(j',l')} \right| \rightarrow f < \infty \end{array} \right\}$$

$$\Sigma_{Fct_q} = \beta\, \beta^T + \Omega^2; (\beta, \Omega) = arg\, min|| \Sigma_{Fct_q} - \Sigma||^2$$
$$\beta \in \Re^{p*q}; \Omega = diag(\omega_1, \ldots, \omega_p); \omega_j > k_0 \in \Re_+$$

with B being and a compact subset of $l_2(N)$, and $\lambda_{min}(\Sigma)$, $\lambda_{max}(\Sigma)$ the smallest and largest eigenvalues of Σ.

The definition of the set Γ_{Fct_q} deserves a few comments.

The condition $\Delta^T \Sigma^{-1} \Delta \geq c^2$ establishes the minimum degree of group separation on Γ_{Fct_q}. For all $\theta \in \Gamma_{Fct_q}$ the optimal misclassification rate is bounded by $1 - \Phi(c/2)$, which becomes a benchmark against which the asymptotic rate of any empirical rule can be compared. Condition $k_1 \leq \lambda_{min}(\Sigma) \leq \lambda_{max}(\Sigma) \leq k_2$ ensures that Σ is always non-singular.

Conditions $\Delta, \gamma \in B$ are necessary technical requirements to ensure that consistent estimation of Δ and γ is possible. We note that when $\Delta \notin l_2(N)$ and $||\Sigma||$ is bounded the expected misclassification rate of the Bayes rule converges to zero when $p \rightarrow \infty$. In that case, it may be possible to find empirical rules with similar perfect asymptotic performance, even if their coefficients diverge from those of the theoretical rule. While such problems may have some interest on their own, they will not be considered here, and we will focus on the more standard conditions where rules approaching perfect group separation are not allowed. Therefore, we assume that $\Delta \in l_2(N)$ and that $\hat{\Delta}, \hat{\gamma}$ are consistent estimators such that $E_\theta||\hat{\Delta} - \Delta||^2 = o(1)$ and $E_\theta||\hat{\gamma} - \gamma||^2 = o(1)$. Known results in the theory of countable Gaussian sequences (see Johnstone 2002 and Lemma 1 in Bickel and

Levina 2004) show that such estimators exist if and only if Δ and γ are restricted to lie on a compact subset of $l_2(N)$.

The previous two conditions are equal (or equivalent) to corresponding conditions assumed by Bickel and Levina (2004) in their theoretical study of the Naive rule.

Conditions

$$\forall j, a \quad \sum_{j',l'} \left| \frac{\partial \beta(j,a)}{\partial \Sigma(j',l')} \right| \rightarrow e < \infty \quad \sum_{j',l'} \left| \frac{\partial \omega_j}{\partial \Sigma(j',l')} \right| \rightarrow f < \infty \tag{12}$$

are new technical requirements, specific to the δ_{Fct_q} rule, that are necessary to ensure that convergence of $\hat{\Sigma_F}$ to Σ translates into convergence of $\hat{\Sigma}_{Fct_q}$ to Σ_{Fct_q}. They imply that for any variable (j) and variable pair (j,l) the contribution of their variances and covariances to the underlying structure of the closest q-factor model can be essentially recovered after a finite number of new variables are added to the model. This seems to be a sensible and reasonable assumption, should it fail no stable q-factor model could be used to approximate the covariance structure defined by the sequence of X variables.

Condition

$$\forall j = 1, \ldots, p \;\; \omega_j > k_0 \in \Re_+ \tag{13}$$

ensures that for $(\Delta, \gamma, \Sigma) \in \Gamma_{Fct_q}, \Sigma_{Fct_q}$ remains always non-singular and well-conditioned. The empirical versions of this condition and formula (8) are central in guaranteeing that, similarly to $\hat{\Sigma}_I$ and unlike $\hat{\Sigma}_F$, $\hat{\Sigma}_{Fct_q}$ can always be inverted and leads to an approximation error $||\hat{\Sigma}^{-1}_{Fct_q} - \Sigma^{-1}_{Fct_q}||$ that can be bounded by a constant times the error $||\hat{\Sigma}_{Fct_q} - \Sigma_{Fct_q}||$.

We are now in condition to state the main result of this section.

Theorem 1. *If* $(ln\ p)/n \rightarrow 0$, *then*

$$lim\ sup_{n\rightarrow\infty}\ \overline{W}_{\Gamma_{Fctq}}(\delta_{Fct_q}) \leq 1 - \Phi\left(\frac{\sqrt{K_{0F_q}}}{1 + K_{0F_q}}\ c\right)$$

$$K_{0F_q} = max_{\Gamma_{Fctq}} \frac{\lambda_{max}(\Sigma_{0Fct_q})}{\lambda_{min}(\Sigma_{0Fct_q})};\ \Sigma_{0Fct_q} = \Sigma^{-\frac{1}{2}}_{Fct_q}\ \Sigma\ \left(\Sigma^{-\frac{1}{2}}_{Fct_q}\right)^T$$

For a proof see Duarte Silva (2009).

We note that the bound defined in Theorem 1 has the same form as the limit found in Bickel and Levina (2004) for the min-max expected error of the Naive rule, replacing the bound (K_0) on the ratio for the eigenvalues of the correlation matrix by K_{0F_q}. This constant reflects the maximum distance between the true covariance and a covariance compatible with the postulated q-factor model. When the data generating process satisfies model (7), Σ_{0Fct_q} is an identity, $K_{0F_q} = 1$, and rule δ_{Fct_q} is asymptotically equivalent to the theoretical Bayes. On the other hand, when K_{0F_q} increases without limit as p grows, the true generating process diverges from

the postulated model and δ_{Fct_q} is asymptotically no better than random guessing. For intermediate cases, with $K_{0F_q} > 1$ but bounded by some finite constant, the performance of δ_{Fct_q}, although not converging to that of the optimal rule, can be close, particularly if K_{0F_q} is never too large.

The main motivation for our proposal, is the fact that K_{0F_q} can be much smaller than K_0 if the true data generating process implies a correlation structure that is far from total independence but close to a structure compatible with a q-dimensional factor model. In such case, Theorem 1 shows that, as p grows, δ_{Fct_q} can approach a considerably smaller expected error rate than δ_I. The promising simulation results presented in the next section suggest that for these conditions, δ_{Fct_q} can still perform much better than δ_I, and δ_F, even for moderate values of p and n.

4 Simulation Study

In order to evaluate the performance of δ_{Fct_q} in finite samples we performed a small simulation experiment with the following design.

We considered balanced samples with two combinations of number of variables and sample size, ($p = 100, n = 200$), and ($p = 100, n = 50$). The first condition intends to illustrate a more traditional situation where the ratio n/p, although relatively small, is still larger than 1, while the second condition starts to explore conditions with $p > n$. For each combination of n and p we considered the following five data generating processes:

Condition A – All variables are independent.
Conditions B, C, D – Variables are generated according to model (7) with $q = 1$(Condition B), $q = 20$ (Condition C) and $q = p$ (Condition D).
Condition E – Variables are generated according to a factor model with p factors, without specific variances.

In conditions B, C, D, and E factor loadings were generated randomly according to an uniform U(0,1) distribution, and then normalized in order to achieve a prespecified communality level. This level was set to 0.5 in conditions B, C and D, while in conditions A and E was respectively equal to 0 and 1. In all conditions we assume that 90% of the variables represent noise and have equal population means (set to 0) in both groups. For the remaining 10% (the signal), we set the means in the second group to the geometric sequence $\mu_1 = (\nu, 0.9\ \nu, 0.9^2\ \nu, \ldots)$ where the constant ν is chosen in order to ensure that the Mahalanobis distance between group centroids is equal to 3. With this setting, the expected rate of the theoretical Bayes rule is equal to $1 - \Phi(1.5) = 0.0668$.

We generated 100 independent training samples, used them to find 100 δ_F, δ_I, and δ_{Fctq} (with $q = 1, 2$ and 3) rules, and evaluated them on an independently generated balanced validation sample with 100,000 observations. Since in this experiment p is not too large, we simply estimated Δ and γ by (2) and (3). However, we note that for problems of higher dimensionality where the influence of

Table 1 Misclassification error rates on the Validation sample: $n = 50$ $p = 100$

	Data conditions				
	A	B	C	D	E
Fisher	0.268	0.297	0.306	0.305	0.500
Naive	0.147	0.368	0.357	0.352	0.500
Fct_{q1}	0.148	0.155	0.177	0.166	0.500
Fct_{q2}	0.149	0.156	0.177	0.167	0.500
Fct_{q3}	0.150	0.157	0.177	0.167	0.500

Table 2 Misclassification error rates on the Validation sample: $n = 200$ $p = 100$

	Data conditions				
	A	B	C	D	E
Fisher	0.173	0.171	0.170	0.168	0.169
Naive	0.090	0.325	0.289	0.278	0.500
Fct_{q1}	0.090	0.090	0.108	0.098	0.500
Fct_{q2}	0.091	0.090	0.108	0.099	0.500
Fct_{q3}	0.091	0.091	0.107	0.099	0.500

noisy variables might be more serious, alternative estimators employing some form of shrinkage or variable selection may be required.

The average misclassification rates in the validation sample are shown in Tables 1 and 2.

We can see in Table 1 that for data condition E with $n = p/2$, none of the methods tried performed better than simple random guessing. However, when $n = 2p$ the Fisher rule performed reasonably well (see Table 2) for this condition . This is a condition that was chosen exactly because of its inherent difficultly and we conjecture that similar structures might not be common in real life applications.

On the other hand, both the δ_I and $\delta_{f_{ctq}}$ rules are quite effective when the data is indeed independent (data condition A). However, the performance of the Naive worsens considerably for all conditions with correlated variables. The choice of the dimensionality (q) assumed by the $\delta_{f_{ctq}}$ rules had only a negligible impact with all three variants tested leading to almost identical results.

The most interesting results are those concerning data conditions C and D. In these conditions, that we believe to be the more realistic ones, each variable has a variability explained in part by a common factor structure and in part by its own characteristics. In both conditions, the true intrinsic dimensionality of the underlying model is considerably higher than the one assumed by the δ_{Fctq} rules, although in condition C (but not in condition D) is smaller than the total number of variables. In both cases, the δ_{Fctq} rules give the best results with an expected error rate that is close the corresponding rate for the condition (B) where the assumed model agreed with the true data generating process. These results are particularly encouraging and we have as top research priority to investigate if they still hold for higher dimensionalities and real data sets.

5 Conclusions and Perspectives

We proposed a new linear discriminant rule capable of incorporating information regarding correlation structures in problems with more variables than observations. Asymptotic properties and moderate sample simulation results suggest that this rule can be quite effective under data conditions where Fisher's and Naive discrimination rules perform poorly.

In the present from, the present rule can be computationally too demanding for very high dimensional problems common in genetic and microarray applications. Variants that try to alleviate the computational burden while retaining some of its desirable statistical properties are currently under investigation. Other avenues of future research include the evaluation of the proposed rule for real-world data sets, and the development of generalizations to more than two groups and quadratic heteroscedastic discrimination problems. An R implementation of the rules described in this paper is available from the author upon request.

References

Bickel, P. J., & Levina, E. (2004). Some theory for Fisher's linear discriminant function, "Naive Bayes" and some alternatives when there are many more variables than observations. *Bernoulli, 10*(6), 989–1010.

Duarte Silva, A. P. (2009). Linear discriminant rules for high-dimensional correlated data: Asymptotic and finite sample results. *Working Papers of the Faculdade de Economia e Gestão*, Universidade Católica Portuguesa at Porto, http://www.porto.ucp.pt/feg/docentes/psilva.

Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discriminant methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association, 97*, 77–87.

Johnstone, I. M. (2002). Function estimation and Gaussian noise sequence models. *Unpublished monograph*, http://www-stat.stanford.edu/~imj.

McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition*. New York: Wiley.

Fast Hierarchical Clustering from the Baire Distance

Pedro Contreras and Fionn Murtagh

Abstract The Baire or longest common prefix ultrametric allows a hierarchy, a multiway tree, or ultrametric topology embedding, to be constructed very efficiently.

The Baire distance is a 1-bounded ultrametric. For high dimensional data, one approach for the use of the Baire distance is to base the hierarchy construction on random projections.

In this paper we use the Baire distance on the Sloan Digital Sky Survey (SDSS, http://www.sdss.org) archive. We are addressing the regression of (high quality, more costly to collect) spectroscopic and (lower quality, more readily available) photometric redshifts. Nonlinear regression is used for mapping photometric and astrometric redshifts.

1 Introduction

In this work we introduce a novel (ultrametric) distance called Baire and show how it can be used to produce clusters through grouping data in "bins". We seek to find inherent hierarchical structure in data, rather than fitting a hierarchy structure to data (as is traditionally used in multivariate data analysis) in an inexpensive computational way.

This paper is structured as follows: firstly we give a definition of the Baire distance; secondly we apply that distance to a chemoinformatics dataset; thirdly we apply the Baire distance to an astronomy dataset; finally we present our conclusions.

P. Contreras (✉)
Department of Computer Science, Royal Holloway, University of London, 57 Egham Hill, Egham TW20 OEX, England
e-mail: pedro@cs.rhul.ac.uk

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_25,

2 Longest Common Prefix or Baire Distance

2.1 Ultrametric Baire Space and Distance

A Baire space consists of countably infinite sequences with a metric defined in terms of the longest common prefix: the longer the common prefix, the closer a pair of sequences. What is of interest to us here is this longest common prefix metric as defined in Murtagh et al. (2008). The longest common prefixes at issue are those of precision of any value. For example, consider two such values, x_{ij} and y_{ij}, which, when the context easily allows it, we will call x and y.

Without loss of generality we take x and y to be bounded by 0 and 1. Each are of some precision, and we take the integer $|K|$ to be the maximum precision. We pad a value with 0s if necessary, so that all values are of the same precision.

Thus we consider ordered sets x_k and y_k for $k \in K$. In line with our notation, we can write x_k and y_k for these numbers, with the set K now ordered. So, $k = 1$ is the first decimal place of precision; $k = 2$ is the second decimal place; . . . ; $k = |K|$ is the $|K|th$ decimal place. The cardinality of the set K is the precision with which a number, x_k, is measured.

Consider as examples $x_k = 0.478$; and $Y_k = 0.472$. In these cases, $|K| = 3$. For $k = 1$, we find $x_k = y_k = 4$. For $k = 2$, $x_k = y_k$. But for $k = 3$, $x_k \neq y_k$.

We now introduce the following distance (case of vectors x and y, with 1 attribute):

$$d_B(x_K, y_K) = \begin{cases} 1 & if\ x_1 \neq y_1 \\ \inf 2^{-n} & x_n = y_n \quad 1 \leq n \leq |K| \end{cases} \tag{1}$$

We call this d_B value Baire distance, which can be shown to be an ultrametric (Murtagh 2004a,b,c, 2005; Murtagh et al. 2008).

Note that the base 2 is given for convenience. When dealing with binary data 2 is the chosen base. When working with real numbers the chosen base is 10.

3 Application to Chemoinformatics

In the 1990s, the Ward minimum variance hierarchical clustering method became the method of choice in the chemoinformatics community due to its hierarchical nature and the quality of the clusters produced. This method reached its limits once the pharmaceutical companies tried processing datasets of more than 500,000 compounds mainly due to its processing requirement of $O(n^2)$.

Datasets of half a million compounds are normal in today's world. There are different ways of encoding a compound to a machine readable form. In chemistry binary fingerprints for chemical compounds are common. The compound is encoded in a fixed length binary string. For details of different encoding systems in chemistry see Brown (2009).

In Murtagh et al. (2008) we applied the Baire distance to a chemoinformatics dataset with the following characteristics:

- 1.2 million chemicals crossed by 1,052 presence/absence attributes (binary matrix)
- The data matrix is highly sparse, occupancy is $\approx$8.6347%
- Chemicals per attribute follow a power law with exponent $\approx$1.23
- Attributes per chemical are approximately Gaussian.

3.1 Dimensionality Reduction by Random Projection

As mentioned above it is a well known fact that traditional clustering methods do not scale well in very high dimensional spaces. A standard and widely used approach when dealing with high dimensionality is to first apply a dimensionality reduction method. For example, Principal Component Analysis (PCA) is a very popular choice to deal with this problem. It uses a linear transformation to form a simplified data set retaining the characteristics of the original data. PCA does this by means of choosing the attributes that best preserve the variance of the data. This is a good solution when the data allows these calculations, but PCA as well as other dimensionality reduction techniques remain expensive, computationally speaking.

In order to apply the Baire distance our first step was to reduce the dimensionality of the original data. We chose to use random projection (Bingham and Mannila 2001; Vempala 2004) not only because of performance but also because of some nice properties of this methodology. Random projection is the finding of a low dimensional embedding of a point set.

In fact random projection here works as a class of hashing function. Hashing is much faster than alternative methods because it avoids the pair-wise comparisons required for partition and classification. If two points (p, q) are close, they will have a very small $\|p - q\|$ (Euclidean metric) value; and they will hash to the same value with high probability. If they are distant, they should collide with small probability.

3.2 Chemoinformatics Data Clustering

In order to cluster the binary data we did the following:

- Normalise the binary data matrix A by column sums; let's call the resulting matrix B
- Produce a random vector Z
- Project B into Z; let's call the resulting matrix R
- Sort the matrix R

– Cluster R applying the longest common prefix or Baire distance; then values that are identical fall in the same cluster.

Following the above process, we show in Murtagh et al. (2008) (p. 728) that for this dataset we can get clusters that are very close to the clusters obtained by k-means. This can be due to many reasons: one reason is that data sparsity in a key factor (i.e. in a large sparse dataset groups are likely to be far from each other, and therefore groups are easier to identify).

4 Application to Astronomy

The Sloan Digital Sky Survey (SDSS) (SDSS 2008) is systematically mapping the sky, producing a detailed image of it and determining the positions and absolute brightnesses of more than 100 million celestial objects. It is also measuring the distance to a million of the nearest galaxies and to one hundred thousand quasars. The acquired data has been openly given to the scientific community.

In this work we are interested into four parameters from a subset of the SDSS data release 5 (Raffaele et al. 2007): declination (DEC), right ascension (RA), spectrometric (Z_{spec}) and photometric (Z_{phot}). In particular we look into redshift data that, for either redshift, vary between 0 and 0.6.

DEC and RA give the position of an astronomical object in the sky. Spectrometric and photometric parameters are two different values obtained to measure the redshift. On one hand we have the spectrometric technique that uses the spectrum of electromagnetic radiation (including visible light) which radiates from stars and other celestial objects. On the other hand we have the photometric technique that uses a faster and more economical way of measuring the redshift, but is less precise than the spectrometric method.

Notice that when talking on the context of speed the advantage of using the Baire metric lies on that it can calculated in $O(n)$ time, unlike many of the traditional clustering method that need a higher computational complexity.

4.1 Clustering SDSS Data Based on a Baire Distance

Figure 1a shows DEC versus RA, i.e. the object's position in the sky. Figure 1b shows the Z_{spec} and Z_{phot} currently used to cluster redshifts. This section of the sky represents approximately 0.5 million coordinate points. As can be observed, various sections of the sky are represented in the data.

Figure 1c–f show graphically how Z_{spec} and Z_{phot} clusters look at different levels of decimal precision. For example, on the one hand we find that values of Z_{spec} and Z_{phot} that have equal precision in the 3rd decimal digit are highly correlated. On the hand when Z_{spec} and Z_{phot} have only the first decimal digit in common correlation is weaker (as shown in Fig. 1e).

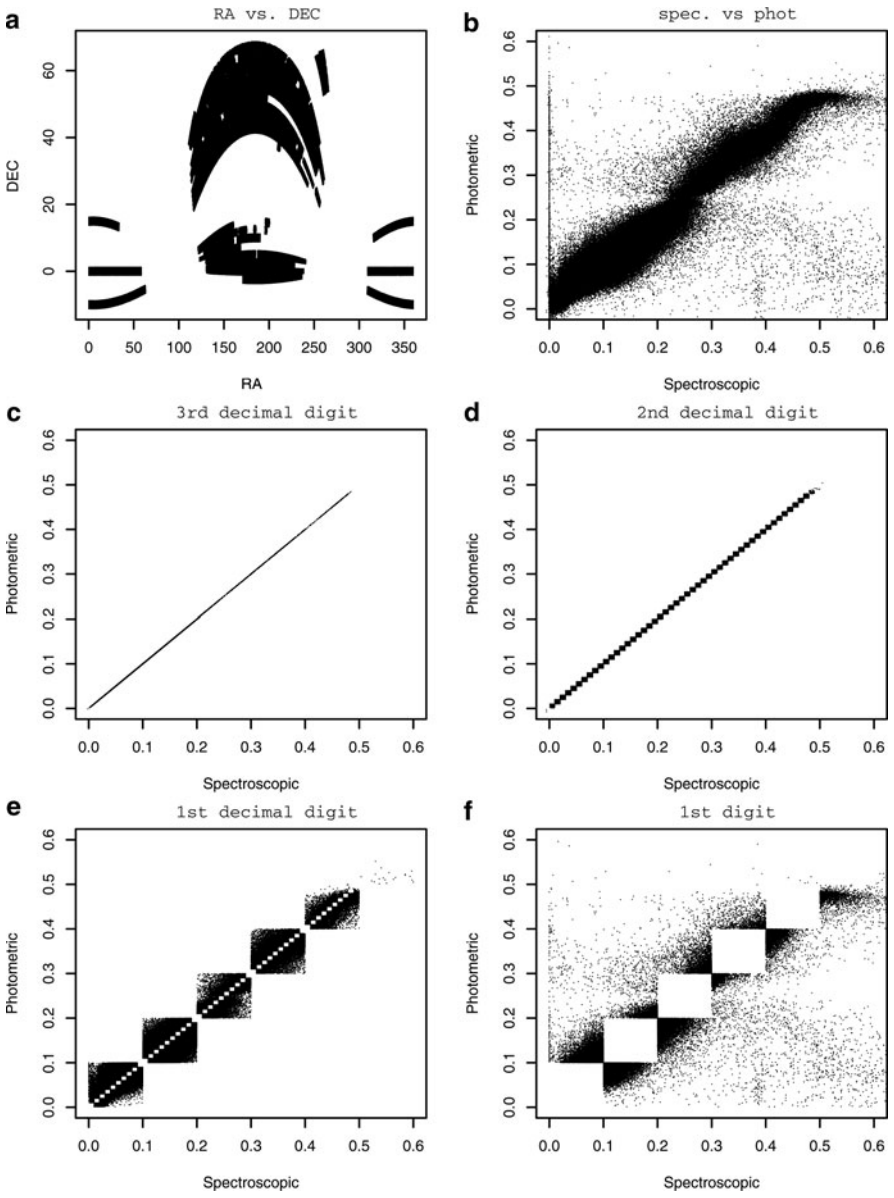


Fig. 1 SDSS data and results for a given precision digit

Notice that in Fig. 1f the data point are scatter around the plot area, these are the data points that have the least information in common, i.e. the data points that do not share any decimal places but the first digit.

Table 1 shows the clusters found for all different levels of precision. In other words this table shows the confidence levels for mapping of Z_{spec} and Z_{phot}. For

Table 1 Clusters based on the longest common prefix

Digit	No	%
1	76.187	17.19
2	270.920	61.14
3	85.999	19.40
4	8.982	2.07
5	912	0.20
6	90	0.02
7	4	–
	443.094	100

example, we can expect that 82.49% of values for Z_{spec} and Z_{phot} to have at least two common prefix digits. Additionally we observe that a considerable number of observations share at least 3 digits in common.

In the following section we a take this notion of clusters even further and compare it to results obtained with the k-means clustering algorithm.

4.2 Baire and K-means Cluster Comparison

In order to establish how "good" the Baire clusters are we can compare them with k-means. Let us recall that our data values are in the interval [0, 0.6[(i.e. including zero values but excluding 0.6). Thus when building the Baire based clusters we will have a root node "0" that includes all the observations. For the Baire distance equal to two we have six nodes (or clusters) with indices "00, 01, 02, 03, 04, 05". For the Baire distance of three we have 60 clusters with indices "000, 001, 002, 003, 004, . . . , 059" (i.e. ten children for each node 00, . . . , 05).

We carried out a number of comparisons for the Baire distance of two and three. For example, we know that for $d_B = 2$ there are six clusters, then we took our data set and applied k-means with six centroids based on the Hartigan and Wong (1979) algorithm. The results can be seen in Table 2, where the columns are the k-means clusters and the rows are the Baire clusters. From the Baire perspective we see that the node 00 has 97084 data points contained within the first k-means cluster and 64950 observations in the fifth. Looking at node 04, all members belong to the cluster 3 of k-means. We can see that the Baire clusters are closely related to the clusters produced by k-means at a given level of resolution.

We can push this procedure further and compare the clusters for d_B defined from 3 digits of precision, and k-means with $k = 60$. Looking at the results from the Baire perspective we find that 27 clusters are overlapping, 9 clusters are empty, and 24 Baire clusters are completely within the boundaries of the ones produces by k-means as presented in Table 3.

Table 2 Cluster comparison based on Baire distance = 2; columns present the k-means clusters (k = 6); rows present Baire nodes

—	1	5	4	6	2	3
00	97084	64950	0	0	0	0
01	0	28382	101433	14878	0	0
02	0	0	0	18184	4459	0
03	0	0	0	0	25309	1132
04	0	0	0	0	0	11116
05	0	0	0	0	0	21

It is seen that the match is consistent even if there are differences due to the different clustering criteria at issue. We have presented results in such as way as to show both consistency and difference.

5 Conclusions

In this work a novel distance called the Baire distance is presented. We show how this distance can be used to generate clusters in a way that is computationally inexpensive when compared with more traditional techniques. This approach therefore makes a good candidate for exploratory data analysis when data sets are very big or in cases where the dimensionality is large. In addition to the advantage of speed, this distance is an ultrametric which can easily be seen as a hierarchy. We applied the Baire definition of distance to two cases:

- In the chemoinformatics case "good" clusters were obtained in the sense that these are close to those produced by k-means.
- In the astronomy case clusters generated with the Baire distance can be useful when calibrating redshifts. In generally, applying the Baire method to cases where digit precision is important can be of relevance, specifically to highlight data "bins" and some of their properties.

Future direction of work includes applying the Baire metric to other data sets. Our particular interest lies in high dimensional and massive data sets like the ones presented in this paper.

Table 3 Cluster comparison based on Baire distance = 3; columns present the k-means clusters (k = 60); rows present Baire nodes

—	21	1	6	38	25	58	32	20	15	13	14	37	17	2	51	4
015	3733	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
004	0	3495	0	0	0	0	0	0	0	0	0	0	0	0	0	0
018	0	0	2161	0	0	0	0	0	0	0	0	0	0	0	0	0
020	0	0	0	1370	0	0	0	0	0	0	0	0	0	0	0	0
001	0	0	0	0	968	0	0	0	0	0	0	0	0	0	0	0
000	0	0	0	0	515	0	0	0	0	0	0	0	0	0	0	0
022	0	0	0	0	0	896	0	0	0	0	0	0	0	0	0	0
034	0	0	0	0	0	0	764	0	0	0	0	0	0	0	0	0
036	0	0	0	0	0	0	0	652	0	0	0	0	0	0	0	0
037	0	0	0	0	0	0	0	508	0	0	0	0	0	0	0	0
026	0	0	0	0	0	0	0	0	555	0	0	0	0	0	0	0
027	0	0	0	0	0	0	0	0	464	0	0	0	0	0	0	0
032	0	0	0	0	0	0	0	0	0	484	0	0	0	0	0	0
030	0	0	0	0	0	0	0	0	0	0	430	0	0	0	0	0
045	0	0	0	0	0	0	0	0	0	0	0	398	0	0	0	0
044	0	0	0	0	0	0	0	0	0	0	0	295	0	0	0	0
039	0	0	0	0	0	0	0	0	0	0	0	0	278	0	0	0
024	0	0	0	0	0	0	0	0	0	0	0	0	0	260	0	0
041	0	0	0	0	0	0	0	0	0	0	0	0	0	0	231	0
042	0	0	0	0	0	0	0	0	0	0	0	0	0	0	225	0
047	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	350
048	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	57
049	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
050	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

References

Bingham, E., & Mannila, H. (2001). Random projection in dimensionality reduction: Applications to image and text data. *KDD '01: Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining. ACM*. San Francisco, California.

Brown, N. (2009). Chemoinformatics – An introduction for computer scientists. *ACM Computing Surveys, 41*(2). Article 8.

Hartigan, J. A., & Wong, M. A. (1979). A K-means clustering algorithm. *Applied Statistics 28*, 100–108.

Murtagh, F. (2004). On ultrametricity, data coding, and computation. *Journal of Classification, 21*, 167–184.

Murtagh, F. (2004). Thinking ultrametrically. In D. Banks, L. House, F. R. McMorris, P. Arabie and W. Gaul (Eds.), *Classification, clustering, and data mining applications* (pp. 3–14). Berlin, Heidelberg, New York: Springer.

Murtagh, F. (2004). Quantifying ultrametricity. J. Antoch (Ed.), *Proceedings in Computational Statistics, Compstat* (pp. 1561–1568). Berlin, Heidelberg, New York: Springer.

Murtagh, F. (2005). Identifying the ultrametricity of time series. *European Physical Journal B., 43*, 573–579.

Murtagh, F., Downs, G., & Contreras, P. (2008). Hierarchical clustering of massive, high dimensional data sets by exploiting ultrametric embedding. Society for Industrial and Applied Mathematics. *SIAM Journal of Scientific Computing, 30*(2), 707–730.

Raffaele, D., Antonino, S., Giuseppe, L., Massimo, B., Maurizio, P., Elisabetta, D., & Roberto, T. (2007). Mining the SDSS archive. I. Photometric Redshifts in the Nearby Universe. ArXiv, arXiv:astro-ph/0703108v2.

SDSS. (2008). Sloan digital sky survey. http://www.sdss.org.

Vempala, S. (2004). *The Random Projection Method* (Vol. 65). DIMACS: Series in Discrete Mathematics and Theoretical Computer Science, Rutgers University. American Mathematical Society.

The Trend Vector Model: Identification and Estimation in SAS

Mark de Rooij and Hsiu-Ting Yu

Abstract Recently, the trend vector model was proposed for the analysis of longitudinal multinomial data. It is a very nice model which graphically represents trends over time for various groups in a low dimensional Euclidean space. The model uses multidimensional scaling tools, which are highly interpretable. The trend vector model, and more general the ideal point classification model has a nasty indeterminacy. De Rooij (2009a,b) solved this problem by using metric multidimensional unfolding with single centering, but this can only be incorporated after the algorithm has converged. Here we show simpler identification results. With the new set of identification constraints the model can be estimated in the SAS software package, which makes the models available to a large audience.

1 Introduction

Longitudinal data arises in many fields of research. When the outcome variable is normally distributed sufficient tools exist for the analysis of such data. For categorical variables the last decennia showed a boost of studies mainly as extensions of generalized linear models. For multinomial unordered categories, i.e. nominal variables, the availability of statistical tools and theory is limited. It can be argued that for nominal outcome variables development is hampered by the dimensionality of the problem. With a discrete outcome variable having G classes the dimensionality is $G-1$, i.e., for each explanatory variable $G-1$ regression parameters have to be estimated and interpreted in a multinomial logit model. To deal with this problem, De Rooij (2009b) proposed the trend vector model, that utilizes multidimensional scaling ideas to reduce the dimensionality.

In the trend vector model the conditional probability $\pi_{gt}(\mathbf{x}_{it})$ of an outcome category g ($g = 1, \ldots, G$) at time point t ($t = 1, \ldots, T_i$), for a subject i ($i = 1, \ldots, n$)

M. de Rooij (✉)
Leiden University Institute for Psychological Research, Leiden, The Netherlands
e-mail: rooijm@fsw.leidenununiv.nl

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_26,

with p-dimensional covariate vector $\mathbf{x}_{it}$ is modeled. The covariate vector contains both time and group information and possibly other explanatory variables. The conditional probability will be modeled by the squared distance between two points in Euclidean space of dimensionality M ($M \leq G-1$): ideal points $\mathbf{y}_{it}$ for the subjects and class-points $\mathbf{z}_g$ for the categories. The ideal points $\mathbf{y}_{it} = (y_{it1}, \ldots, y_{itM})^{\mathsf{T}}$, which are gathered in a matrix $\mathbf{Y} = (\mathbf{y}_{11}, \ldots, \mathbf{y}_{1T_1}, \ldots, \mathbf{y}_{nT_n})^{\mathsf{T}}$, are a linear combination of the predictor variables $\mathbf{X}$, i.e.,

$$\mathbf{Y} = \mathbf{XB},$$

where $\mathbf{X} = (\mathbf{x}_{11}, \ldots, \mathbf{x}_{1T_1}, \ldots, \mathbf{x}_{nT_n})^{\mathsf{T}}$. The conditional probability that subject i at time point t will be in class g is then equal to

$$\pi_{gt}(\mathbf{x}_{it}) = \frac{\exp(-d^2_{(it)(g)})}{\sum_k \exp(-d^2_{(it)(k)})},$$

where $d^2_{(it)(g)}$ is the squared Euclidean distance between the ideal point for subject i at time point t and the class point for category g in M-dimensional space, i.e.,

$$d^2_{(it)(g)} = \sum_{m=1}^{M} (y_{itm} - z_{gm})^2.$$

The trend vector model is estimated by maximizing

$$L = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \log \prod_{g=1}^{G} \pi_{gt}(\mathbf{x}_{it})^{f_{itg}}, \tag{1}$$

which is the likelihood function for cross sectional data. In our case it is not a true likelihood, since the dependencies among the repeated responses are not taken into account. As is shown in Liang and Zeger (1986) maximizing L with repeated measurements does provide consistent estimates of the model parameters. However, standard errors computed from the Hessian or information matrix of this function are generally biased. To deal with this bias, Liang and Zeger (1986) introduced a sandwich estimator. For generalized linear models Liang and Zeger (1986) also adapt the estimation equations using these sandwich functions to obtain generalized estimating equations (GEE). Various forms of correlation structures have been proposed to obtain the sandwich function like independence, exchangeable, first order auto regressive, or unstructured. When maximizing (1) we implicitly use the GEE framework with independence assumptions to estimate the model parameters.

2 Example

We use data published in Adachi (2000) to illustrate our trend vector model. In this research Japanese boys and girls were asked after their preferred type of TV programme at five time points. The five time points are the first year of elementary school (ages 6–7), the fourth year of elementary school (ages 9–10), first year of junior high school (ages 12–13), first year of high school (ages 15–16), and as university freshmen (ages 18–20). The TV programme categories are Animation (A), Cinema (C), Drama (D), Music (M), Sport (S), and Variety (V). Frequencies of preference for each category at the five time points are given in Table 1.

We constructed a time variable T by using the midpoints of the ages at the specific time points as scores (i.e., 6.5 for the first time point, 9.5 for the second, etc.) and center around the mean (12.25). The question is whether boys and girls differ in their trends in TV-viewing behavior and how the trends look like. Figure 1 gives the solution where there is a main effect for gender and the time development follows a polynomial of degree two. The regions where the odds for a give category are highest are given by the regions around the class points. The two arrows give the trends over time for boys and girls, the markers represent the ages 6 till 20. Boys and girls have the same quadratic trend over time, but a different starting position. Although boys and girls both start with preferring 'animation' the trend for boys passes 'variety', 'music', and ends in 'sports', while that for girls passes 'drama' and ends in 'music'.

Table 1 Frequencies of preference for TV programme categories for Japanese boys and girls

Group	Preference	Time point				
		$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$
Boys	A	36	26	2	1	0
	C	0	2	4	4	10
	D	1	0	7	14	4
	M	1	2	10	10	10
	S	3	4	9	7	14
	V	8	15	17	13	11
Girls	A	49	31	8	2	1
	C	0	0	1	3	11
	D	0	6	26	21	12
	M	0	1	6	13	15
	S	0	1	0	2	4
	V	2	12	10	10	8

A stands for Animation, C for Cinema, D for Drama, M for Music, S for Sport, and V for Variety

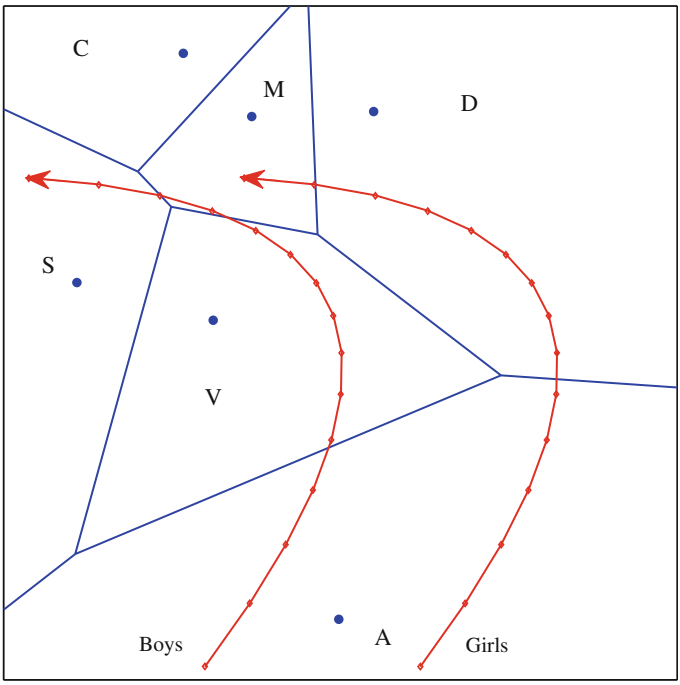


Fig. 1 Solution of the trend vector model on the TV preference data. The straight lines represent boundaries where the odds for two categories are even. The two curves represent the trends for boys and girls, with markers for ages from 6 till 20

3 Identification Problems

The parameters of the trend vector model are the regression weights **B** and the class points **Z**. The model has rotational freedom and a more intricate indeterminacy, that is, the probabilities remain the same when a constant is added for each subject:

$$\pi_{gt}(\mathbf{x}_{it}) = \frac{\exp(-d^2_{(it)(g)})}{\sum_k \exp(-d^2_{(it)(k)})} = \frac{\exp(-d^2_{(it)(g)} + c_i)}{\sum_k \exp(-d^2_{(it)(k)} + c_i)}.$$

So, we can add a constant to each subjects' squared distances to the class points without changing the probabilities. We call this second type of indeterminacy the 'multinomial indeterminacy'. De Rooij (2009a) shows that these indeterminacies allow for the transformations of **Z** and **B** to

$$\mathbf{Z}_* = \mathbf{1}\mathbf{v}^T + \mathbf{Z}\mathbf{T}$$
$$\mathbf{B}_* = \mathbf{B}(\mathbf{T}^{-1})^T$$

with $\mathbf{T}$ an $M \times M$ matrix and $\mathbf{v}$ an $M \times 1$ vector, under the restriction that $\text{diag}(\mathbf{Z}_*\mathbf{Z}_*^T) = \text{diag}(\mathbf{Z}\mathbf{Z}^T) + q\mathbf{1}$ for any q. The total number of indeterminacies equals $\max(M(M-1)/2, M(M+1)-(G-1))$, as is shown in detail in De Rooij (2009a).

3.1 De Rooij's Solution

In order to obtain an identified solution De Rooij (2009a) observed that row-wise centering makes unidentified solutions equal. Let $\mathbf{\Pi} = \{\pi_{gt}(\mathbf{x}_{it})\}$, $\mathbf{\Delta} = \log \mathbf{\Pi}$, and $\mathbf{J} = \mathbf{I}_G - \mathbf{1}_G\mathbf{1}_G^\intercal/G$. Then it was noted that $-\mathbf{\Delta}\mathbf{J} = \mathbf{D}\mathbf{J}$, with $\mathbf{D}$ the matrix with squared Euclidean distances between ideal points and class points for any unidentified solution. This makes it possible to use metric unfolding with single centering for identification. This procedure works fine, except in the situation of maximum dimensionality, i.e. $M = G - 1$. In this case de Rooij identified the solution by a transformation of $\mathbf{Y}$ such that $\mathbf{Y}^\intercal\mathbf{Y} = n\mathbf{I}$, which can be obtained using a singular value decomposition, and solving for the class points (see De Rooij 2009a for details). This identification solution can only be implemented after the algorithm has converged, i.e. an extra analysis step is needed. It would be much nicer if direct constraint can be placed on the configuration such that it is identified from the beginning. Here we will propose a set of constraints that can be imposed at the start of the optimization, in fact it are fixed coordinates constraints.

3.2 Simpler Solution

Looking at the number of indeterminacies $\max(M(M-1)/2, M(M+1)-(G-1))$ we see that the first term is related to rotation of the space while the second term amount to indeterminacies implied by the multinomial distribution (2). The first indeterminacy is solved, like in De Rooij (2009a) by setting the upper triangular part of $\mathbf{B}$ equal to zero. The other indeterminacies are solved by restrictions on the class points $\mathbf{Z}$. In all cases these restrictions amount to centering and scaling of the class points. Note that the origin of the space is determined through the matrix with explanatory variables $\mathbf{X}$, since this matrix does not include a constant vector, and thus when $\mathbf{x}_{it} = \mathbf{0}$ then $\mathbf{y}_{it}$ is $\mathbf{0}$.

In the following we explicitly deal with two dimensional solutions, but the details will make transfers to higher dimensional solutions clear. The term $M(M+1) - (G-1))$ with $M = 2$ shows that for $G = 3, 4, 5$ indeterminacies exists, while for $G > 5$ only rotational indeterminacies remain, which are solved as described above.

The case of $G = 3$

In this case, we have to impose three additional restrictions on the class points. This can be accomplished by setting the 'scale' on both dimensions and centering on

one of the dimensions. Therefore, restrict $z_{11} = 1$ and $z_{22} = 1$, i.e. the scale on both dimensions is fixed now, since the origin was already determined. Moreover, let $z_{32} = -(z_{12} + z_{22})$ to center along the second dimension.

The case of $G = 4$

In this case, we have to impose two restrictions on the class points. Compared to the three class case we can drop the centering condition, but we have to keep the scaling restrictions. An identified solution is obtained by setting $z_{11} = 1$ and $z_{22} = 1$.

The case of $G = 5$

In this case, we have to impose a single restriction on the class points. Compared to the four class case we can drop one scaling condition, i.e. only a scaling condition on the first dimension has to be imposed. An identified solution is obtained by setting $z_{11} = 1$.

4 Estimation with SAS proc nlmixed

It is not well known that the `NLMIXED` procedure of `SAS` has a general optimization tool in it, and therefore it can be used for models without random effects. Detailed information about the `NLMIXED` procedure can be found on http://support.sas.com/documentation/cdl/en/statug/59654/HTML/default/nlmixed_toc.htm. Here we provide the code for a model with three classes and four predictor variables.

```
001 proc nlmixed data=trenddata;
002 PARMS
003 b1_1-b1_4=0 b2_2-b2_4=0
004 z1_2 = 0 z2_1 = 0 z1_3=0;
005 /*Code linear predictors */
006 y1 = b1_1*x1 +b1_2*x2 + b1_3*x3 + b1_4 *x4;
007 y2 = b2_2*x2+ b2_3*x3 + b2_4*x4;
008 /* identification constraints */
009 z1_1 = 1; z2_2=1; /*scaling*/
010 z2_3 = -(z2_1+z2_2); /*centering second dim*/
011 /*Code squared distances*/
012 dist1= (y1-z1_1)*(y1-z1_1) + (y2-z2_1)*(y2-z2_1);
013 dist2= (y1-z1_2)*(y1-z1_2) + (y2-z2_2)*(y2-z2_2);
014 dist3= (y1-z1_3)*(y1-z1_3) + (y2-z2_3)*(y2-z2_3);
```

```
015 /*probabilities*/
016 denom = exp(-(dist1))+ exp(-(dist2))+ exp(-(dist3));
017 if (resp = 1) then p = exp(-(dist1)) / denom;
018 else if (resp = 2) then p = exp(-(dist2)) / denom;
019 else if (resp = 3) then p = exp(-(dist3)) / denom;
020 /*Define likelihood*/
021 if (p > 1e-8) then ll = log(p);
022 else ll = -1e100;
023 model resp   general(ll);
024 run;
```

In lines 002–004 starting values for the parameters are declared, often starting with values for the regression weights equal to zero and using dispersed class points works well. One has to be aware of the fact that the algorithm may run into local optima, so that multiple start points have to be performed. In lines 005–007 the linear combinations that define the ideal points are given. Identification constraints for this three class problem, as discussed above, are implied in lines 008–010. Then for each response category the squared distance from an ideal point till the class points are given in lines 011–014. From these distances conditional probabilities are obtained in lines 015–019. Finally, the optimization function is defined in lines 020–023.

As is shown in Yu and de Rooij (2010), likelihood ratio statistics and BIC statistics are favored for model selection. The BIC is given by

$$\text{BIC} = -2 * L + \log(N) \times \text{npar}$$

where npar denotes the number of parameters. It should be noted that the BIC value that SAS is giving is not correct, since it uses in the penalty term $N = \sum_{i=1}^{n} T_i$, where T_i is the number of observations for subject i and it should be the total number of subjects ($N = n$). This is easily repaired by hand, since the deviance and the number of parameters are given in the output file. The standard errors that are reported by SAS are biased, since we have no correct likelihood function, i.e., these should not be used for model selection purposes.

5 Conclusion

We provided a simpler set of identification constraints for the trend vector model and gave SAS code in order to fit the model. We think this is valuable, since SAS is a computer program that many people have access to.

References

Adachi, K. (2000). Scaling of a longitudinal variable with time-varying representation of individuals. *British Journal of Mathematical and Statistical Psychology, 53*, 233–253.

De Rooij, M. (2009a). Ideal point discriminant analysis revisited with an emphasis on visualization. *Psychometrika, 74*, 317–330.

De Rooij, M. (2009b). Trend vector models for the analysis of change in continuous time for multiple groups. *Computational Statistics and Data Analysis, 53*, 3209–3216.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*, 13–22.

Yu, H.-T., & de Rooij, M. (2010). Model selection for trend vector models with longitudinal multinomial outcomes *Submitted paper*.

Discrete Beta-Type Models

Antonio Punzo

Abstract A more interpretable parameterization of a beta density is the starting point to propose an analogous discrete beta ($d.b.$) distribution assuming values on a finite set. Thus a smooth estimator using $d.b.$ kernels is considered. By construction, it is both well-defined and free of boundary bias. Taking advantage of the discrete nature of the data, a technique of smoothing parameter selection is also proposed in moderate-to-large samples. Finally, a real data set is analyzed in order to appreciate the advantages of this nonparametric proposal.

1 Introduction

Let X be a discrete random variable ($r.v.$), assuming values on the finite set $\mathcal{S} = \{0, 1, \ldots, k\}$. Moreover, let $p(x) = P(X = x)$, $x \in \mathcal{S}$, be the unknown probability mass function ($p.m.f.$) of X.

There exists a wide variety of practical problems in which the phenomenon of interest is inherently described by X. For example, in item response theory context, X could be either the number of correct responses in a test with k nominal items, or the so-called raw total score in a test with I *ordinal* items having response categories $0, 1, \ldots, M$ (with these conditions, the extreme raw total scores are 0 and $k = I \cdot M$). Many other phenomena, usually described by a *ratio-continuous* variable, could be analyzed by X too. The distinction between "discrete" and "continuous" is indeed only *conceptual* because of both boundedness of the usual measuring instruments and conventional discretization. For example, the age of a subject, that is by nature a continuous variable, should be analyzed by X since every time it belongs to the interval $[j, j + 1)$, the convention suggests attributing the value j, $j \in \mathcal{S}$. Consequently any variable, even in principle continuous, is pragmatically discrete. Again, as underlined in Punzo and Zini (2008), all practical problems in which the phenomenon of interest assumes values on the compact

A. Punzo
Dipartimento di Economia e Metodi Quantitativi, Università di Catania, Italy
e-mail: antonio.punzo@unict.it

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_27,

interval $[0, 1]$ (*e.g.*, rates, positive normalized indexes, and so on) can be described by X after a convenient discretization; in these cases, a choice of k equal to 100, 1,000, or 10,000, permits a simple interpretation of the x values in %, ‰ and ‱ terms, respectively (in truth, if the observed values have c decimal places, $k = 10^c$ guarantees the maximum degree of approximation).

Thus, a statistical model for exploring and presenting the distribution of the data becomes probably more important in the discrete case than in the continuous one. In large samples, it is natural to use the naive unbiased estimator of $p(x)$, that is, its empirical counterpart $f_x = n_x/n$, being n_x the absolute frequency of observations equals to x in the sample $x_1, \ldots, x_n$; obviously, $n = \sum_{x=0}^{k} n_x$. However, in small-to-moderate samples, such an estimator is not so compelling. Moreover, it may be too "rough". In such circumstances at least two routes are possible. First, to rely on some assessed simple parametric structure for $p(x)$ (*e.g.* a binomial *p.m.f.*); second, according to the general philosophy that, at the varying of $x \in \mathcal{S}$, f_x gives a version of $p(x)$ obscured by noise and that judicious smoothing can reduce this noise without distorting the true picture, to attempt to smooth the $k + 1$ naive estimates. Naturally, the choice of one route does not exclude the other: for example, if no *a priori* information about the unknown *p.m.f.* is available, a preliminary smoothing analysis could give valuable indication of features such as skewness and multimodality, useful in suggesting simple parametric formulations or, *vice versa*, in rejecting any parametric specification.

The paper, making a contribution in both the above-mentioned directions, is structured as follows. From a parametric viewpoint, in Sect. 2, a discrete analogue defined on $\mathcal{S}$ of a beta density is proposed, starting from one of its more interpretable parameterizations. This discrete analogue is used in Sect. 3, from a nonparametric viewpoint, to define a kernel smooth estimator of $p(x)$ having as kernels discrete beta distributions opportunely "placed" in the single observations. In order to select the smoothing parameter of this model, a technique taking advantage of the discrete nature of the data is proposed in Sect. 3.1 for moderate-to-large samples. Finally, in Sect. 4, a real data set is analyzed in order to appreciate advantages and motivations of the nonparametric proposal.

2 A Re-parameterized Discrete Beta Distribution

In this section, in order to obtain a more flexible parametric model for $p(x)$ than the existent ones, a discrete version of a beta distribution is proposed. To do this, a more interpretable parameterization of a beta distribution will be taken into account. In detail, let

$$f(y; k, \varepsilon, m, h) = \frac{(y+\varepsilon)^{\frac{m+\varepsilon}{h(k+2\varepsilon)}}(k+\varepsilon-y)^{\frac{k+\varepsilon-m}{h(k+2\varepsilon)}}}{(k+2\varepsilon)^{\frac{h+1}{h}} B\left[\frac{m+\varepsilon}{h(k+2\varepsilon)}+1, \frac{k+\varepsilon-m}{h(k+2\varepsilon)}+1\right]},$$

with,

$$-\varepsilon \leq y \leq k+\varepsilon, \tag{1}$$

with $m, h \in \mathbb{R}$ and $\varepsilon > 0$, be an alternative parameterization of a beta distribution defined on $[-\varepsilon, k+\varepsilon]$. From the standard theory on the beta distribution, a *r.v.* Y with density function (1) has variance

$$\mathsf{Var}(Y) = \frac{h\left[(m+\varepsilon)+h(k+2\varepsilon)\right]\left[(k+\varepsilon-m)+h(k+2\varepsilon)\right]}{(2h+1)^2(3h+1)}. \tag{2}$$

The advantage of the re-parameterization (1) is in terms of graphical interpretation of the parameters m and h. Specifically, a point of maximum (minimum) in correspondence to $y = m$ is obtained if $m \in [-\varepsilon, k+\varepsilon]$ and $h > 0$ ($h < 0$). Moreover, considering $h > 0$ and $m \in [-\varepsilon, k+\varepsilon]$ in (2), the variance results directly proportional to the value of h: in detail, the limit of (2), as $h \to 0^+$, is zero, while as h becomes large the limit is $(k+2\varepsilon)^2/12$, that is the variance of a uniform distribution defined on $[-\varepsilon, k+\varepsilon]$ (note that $f(y; k, \varepsilon, m, h)$ converges to a uniform distribution when $h \to \infty$).

Discretizing (1) on $\mathcal{S}$, the following discrete beta (*d.b.*) distribution can be considered

$$p(x; k, \varepsilon, m, h) = \frac{(x+\varepsilon)^{\frac{m+\varepsilon}{h(k+2\varepsilon)}}(k+\varepsilon-x)^{\frac{k+\varepsilon-m}{h(k+2\varepsilon)}}}{\sum_{j=0}^{k}(j+\varepsilon)^{\frac{m+\varepsilon}{h(k+2\varepsilon)}}(k+\varepsilon-j)^{\frac{k+\varepsilon-m}{h(k+2\varepsilon)}}}, \quad x \in \mathcal{S}, \tag{3}$$

with $m, h \in \mathbb{R}$ and, as before, $\varepsilon > 0$ (for a natural parameterization of a *d.b.* distribution, see Punzo and Zini 2008). The considerations on the interpretation of the parameters m and h in (1) can be re-stated as follows: if $m \in \mathcal{S}$, then $p(x; k, \varepsilon, m, h)$, at the varying of $x \in \mathcal{S}$, represents a *p.m.f.* with a single mode in $x = m$ and, in addition when $h \to 0^+$, it tends to a Dirac delta function in $x = m$. Conversely, if $h \to \pm\infty$, then $p(x; k, \varepsilon, m, h)$ tends to a discrete uniform distribution. The effect of varying h, the other parameters being fixed, is illustrated in Fig. 1, while the graphical effect of varying m, fixed h, k and ε, is displayed in Fig. 2. Note that in Fig. 2 the variation of m is stopped on the middle point $x = 50$ since two distributions of kind (3), having modes in m and $k-m$, other parameters being equal, are each other's reflection in $k/2$, that is

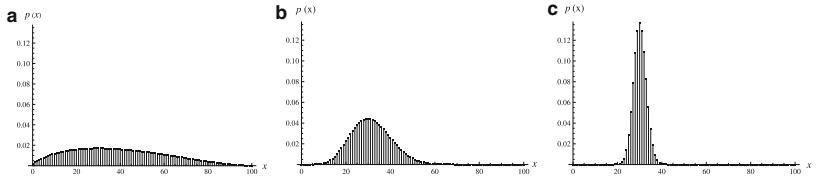


Fig. 1 The effect of varying h ($k = 100$, $\varepsilon = 0.5$, $m = 30$)

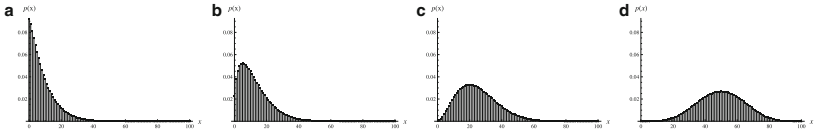


Fig. 2 The effect of varying m ($k = 100$, $\varepsilon = 0.5$, $h = 0.1$)

$$p(x; k, \varepsilon, m, h) = p(k - x; k, \varepsilon, k - m, h). \tag{4}$$

It is interesting to note that the value $\varepsilon = 1/2$ can be considered as a sort of *continuity correction*. Computational efforts indeed shown that, if $h > 0$ (in order to avoid U-shapes) and $m \in \mathcal{S}$, then the individual probabilities (3) are related to the beta distribution (1) by the following approximation

$$p\left(x; k, \frac{1}{2}, m, h\right) \approx \int_{x-\frac{1}{2}}^{x+\frac{1}{2}} f\left(y; k, \frac{1}{2}, m, h\right) dy, \quad x \in \mathcal{S}, \tag{5}$$

in which the error committed is almost always negligible and it is even null when $h \to \infty$. The left-hand side of (5) equals the area of a rectangle with height $p(x; k, 1/2, m, h)$ whose basis is the interval of unitary length centered at x, $x \in \mathcal{S}$. Note that the degree of approximation decreases when at least one of the following conditions holds: $h \to 0^+$, m is located near to the boundary, k is small.

3 Smoothing by Discrete Beta Kernels

Kernel smooth estimators for $p.m.f$s have been considered by several authors (see, *e.g.*, Aitchison and Aitken 1976; Titterington 1980; Wang and van Ryzin 1981; Hall and Titterington 1987). These estimators are merely a sum of n (usually symmetric) "bumps" (the so-called kernels), with equal weights $1/n$, placed at the observations by means of their mode. Unfortunately, as stressed in Chen (1999) for the continuous case, while using a symmetric kernel is appropriate for fitting distributions with unbounded supports, it is not adequate for distributions with compact or bounded from one end only supports as it causes boundary bias. The cause of boundary bias is due to the fixed symmetric kernel which allocates weight outside the support when smoothing is made near the boundary. This section shows how a convenient use of $d.b.$ kernels automatically permits a solution to this problem. Moreover, the resulting estimator produces well-defined estimates, that is, estimates satisfying all the fundamental properties of a $p.m.f$.

Consider a $d.b.$ distribution with $m \in \mathcal{S}$, $h > 0$, and $\varepsilon = 1/2$. Placing it in correspondence with each single observation by putting $m = x_i$ in (3), it is possible to consider the following kernel smooth probability estimator

$$\widehat{p}(x;h) = \frac{1}{n}\sum_{i=1}^{n} p\left(x;k,\varepsilon=\frac{1}{2},m=x_i,h\right) = \frac{1}{n}\sum_{i=1}^{n} k_h(x;x_i), \quad x \in \mathcal{S}, \quad (6)$$

where $k_h(x;x_i)$ and h are the $d.b.$ kernel and the *smoothing parameter*, respectively. By construction, (6) defines a $p.m.f.$ The extension of this model to the case $\varepsilon \neq 1/2$ is straightforward. An equivalent representation of (6) is

$$\widehat{p}(x;h) = \sum_{m=0}^{k} f_m k_h(x;m), \quad x \in \mathcal{S}, \quad (7)$$

that is merely a mixture, with "observed" weights f_m, of $k+1$ $d.b.$ components $k_h(x;m)$, of equal parameter h, with mode in m, $m \in \mathcal{S}$.

Two quantities characterize the nonparametric estimator (6)–(7): the smoothing parameter h and the $d.b.$ kernels $k_h(x;x_i)$. The former can be considered as smoothing parameter for the following considerations: according to the results of Sect. 2, if h is chosen too large, all details, modes, spurious or otherwise, may be obscured; *vice versa*, as h becomes small, spurious fine structure becomes visible. The limit as $h \to 0^+$ is a sum of Dirac delta function spikes at the observations; consequently, $\widehat{p}(x;h)$ converges to the empirical frequency distribution f_x, $x \in \mathcal{S}$. As regards the $d.b.$ kernels, they obey the fundamental graphical properties of a kernel function. In detail, they are non-negative, sum to one, assume their maximum value when $x = x_i$, and are smoothly non-increasing as $|x - x_i|$ increases. The only unconventional property is their skewness: indeed, fixed $h > 0$, the kernel shape changes naturally according to the position where the observation x_i falls (see Fig. 2); however, based on (4), equal weight (equal kernel) at the symmetrical observations falling in x and $k - x$ is attributed. This characteristic, along with the fact that the support $\mathcal{S}$ of a $d.b.$ kernel matches the support of the unknown $p.m.f.$, constitutes a natural remedy to the problem of boundary bias.

Finally, although the $d.b.$ kernel estimator (6) is in philosophy similar to the beta kernel density estimator proposed by Chen (1999), there are differences both in context (Chen's estimator is defined on $[0,1]$) and in structure (the way the beta kernels are adopted). Because of the structural difference, unlike the probability estimator in (6), Chen's estimator produces a $p.m.f.$ non-integrating to one.

3.1 Choosing the Smoothing Parameter h

In order to solve the problem of choosing the smoothing parameter, it should never be forgotten that the appropriate choice of h will always be influenced by the purpose for which the $p.m.f.$ estimate is to be used. Taking advantage of the discrete nature of the data, a method permitting the choice of h subjectively on the basis of the degree of fit with the observed distribution that the user considers right, will be here proposed. In effect, as pointed out in Titterington (1985), it would

seem unreasonable to choose h in such a way that the resulting $\widehat{p}(x;h)$ provides an unacceptably poor fit to the observed data.

Considering the $k+1$ observed frequencies in (7) as *a priori* fixed, it is possible to use (as function of h) the well-known Pearson's chi-squared statistic

$$X^2(h) = \sum_{x=0}^{k} \frac{[n_x - \widehat{n}_x(h)]^2}{\widehat{n}_x(h)},$$

where $\widehat{n}_x(h) = n\widehat{p}(x;h)$, having approximately the $\chi^2_{(k-1)}$ distribution when $n/k \to \infty$. Logically, the value $X^2(h) = 0$, corresponding to a perfect fit between the model and the empirical frequency distribution, can be obtained when $h \to 0^+$. Now, let $\chi^2_{(k-1;\alpha)}$ be the quantile of a $\chi^2_{(k-1)}$ having on the right a probability mass equal to $1-\alpha$. Choosing in advance the degree of fit $1-\alpha$ that is retained opportune, the value of h such that $X^2(h) = \chi^2_{(k-1;\alpha)}$ can be selected. In order to do this, numerical procedures could be used to solve, for example, the following problem:

$$\widehat{h} = \arg\min_{h>0} \left[X^2(h) - \chi^2_{(k-1;\alpha)} \right]^2. \tag{8}$$

4 Application to a Real Data Set

With the aim of underlining the advantages of the $d.b.$ kernel estimator, a real data set concerning the number of deaths – subdivided both by year of death and age of death for individual – that occurred in Catania during the period 2001–2004, has been analyzed. The $n = 13{,}592$ data utilized here were made available by the Data Archive at the office of vital statistics of Catania.

In detail, let $X :=$ "age of death for individual" be the variable of interest. Certain relevant summary statistics are given in Table 1. Moreover, the mortality barplots, placed one behind the other in chronological order, are depicted in Fig. 3(a). Not surprisingly, also confirming the informations in Table 1, the barplots are skewed with a short tail for older ages. Again, a strong incidence of mortality in the first year of life, as well as its decrease over time, is made strikingly clear. Not as clear, due to both the ragged behavior of the barplots and the complexity of the phenomenon, are other general features of interest as, for example, the most common age of death, the locations of the modes, changing in the structure over time, and so on, all information that could have interesting practical interpretations.

Additional information can be gleaned by examining Fig. 3(b) where the $d.b.$ beta kernel estimator (6)–(7) is fitted to the observed data over time. Observe that, for each considered year, k has been posed equal to the maximum observed age of death for individual (see Table 1) in order to avoid null observed frequencies. Moreover, having the data in Fig. 3(a) an evident complex structure, the smoothing parameter h for each year, estimated minimizing in `Mathematica` environment

Table 1 Summary statistics

year X	2001	2002	2003	2004	2001–2004
$\max(X)$	103	107	106	104	107
$\mu(X)$	75.6953	76.5704	77.3759	77.3582	76.7099
$\sigma(X)$	15.5484	15.0517	14.1953	14.2074	14.8143
Skew(X)	−1.9236	−1.9469	−1.8621	−1.8128	−1.9058

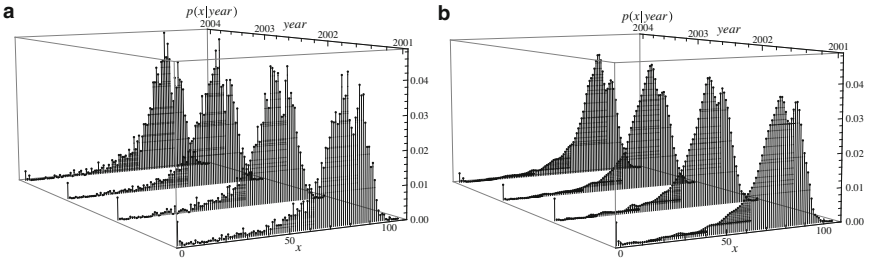


Fig. 3 Probability mass function estimators over time (from 2001 to 2004)

the quantity (8), has been chosen to guarantee a 99% degree of fit. The estimated values of h, as well as the graphical representation of the objective function in (8) plotted against h, is displayed in Fig. 4.

Fig. 3(b) indicates that these populations each have at least two modes – probably placed in the points of most common age of death in the subpopulations of males (first mode) and females (second mode) – and that there is some change in the structure over time: more "deaths belonging to the second mode" enter in the first so that the first peak becomes slowly more dominant and the second less prominent. Surprisingly, the information on the impact of mortality in the first year of life, as well as its decrease over time, is also preserved with respect to the barplots in Fig. 3(a). Careful examination reveals another general feature of interest: some small but prominent bumps in the "curves", between the ages of 15 and 50, are visible too. This "excess mortality" is probably due to an increase in a variety of risky activities, the most notable being obtaining a driver's license.

In line with the considerations in Sect. 1, Fig. 3(b) suggests rejecting any discrete parametric specification since the existent models of this category, including among these the $d.b.$ distribution, can only produce unimodal distributions *per se*. Again, the existence of a group-structure for sex in the data could suggest, for example, a finite mixture of two $d.b.$ distributions with modes in correspondence of the most common ages of death for males and females and weights related to the number of deaths for sex: unfortunately, also in this case, other fundamental characteristics such as, for example, the strong impact of mortality in the first year of life, should be oversmoothed. In conclusion, the complexity of the phenomenon needs a statistical model flexible enough to capture the characteristics of the mortality distribution;

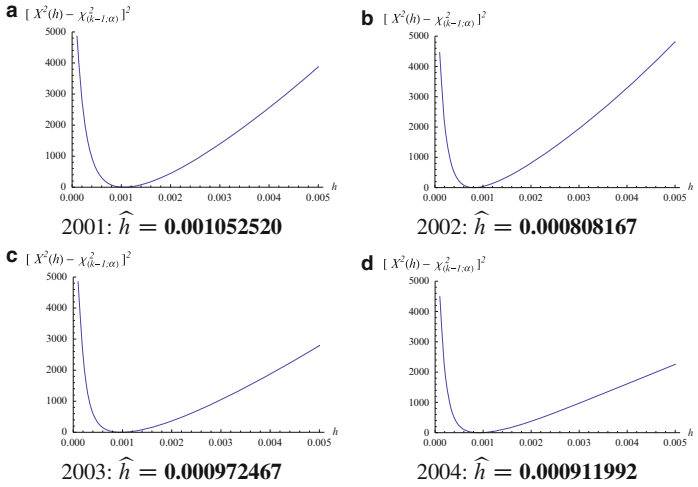


Fig. 4 Plot of $\left[X^2(h) - \chi^2_{(k-1;\alpha)}\right]^2$ as a function of h, fixed $\alpha = 0.01$

in these terms, the $d.b.$ kernel estimator provides an example of how a statistical model can show structures that are very difficult to see by classical methods.

5 Concluding Remarks

In this paper the problem of modeling discrete distributions defined on a finite support has been faced making both a parametric and a nonparametric contribution. The building block is the discrete beta ($d.b.$) distribution obtained by simple discretization of a beta density conveniently parameterized. Kernel estimators using $d.b.$ kernels are also considered. By construction they are free of boundary bias, well-defined (non-negative and summing to one), and easy both in concept and in implementation. Moreover, it is remarked that this smooth estimator makes it possible to estimate discrete distributions of greater complexity with the degree of fit that is viewed as convenient. As a final remark, this article leaves an important open question: the computation of the moments of the $d.b.$ distribution in terms of its parameters. In order to do this, an in-depth theoretical study of the approximation suggested in (5) could be a useful starting point.

References

Aitchison, J., & Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika, 63*(3), 413–420.

Chen, S. X. (1999). Beta kernel estimators for density functions. *Computational Statistics & Data Analysis, 31*(2), 131–145.

Hall, P., & Titterington, D. M. (1987). On smoothing sparse multinomial data. *Australian & New Zealand Journal of Statistics, 29*(1), 19–37.

Punzo, A., & Zini, A. (2008). *Discrete approximations of continuous and mixed measures on a closed interval.* Technical Report 160, Università di Milano-Bicocca, Dipartimento di Metodi Quantitativi per le Scienze Economiche e Aziendali.

Titterington, D. M. (1980). A comparative study of kernel-based density estimates for categorical data. *Technometrics, 22*(2), 259–268.

Titterington, D. M. (1985). Common structure of smoothing techniques in statistics. *International Statistical review, 53*(2), 141–170.

Wang, M., & van Ryzin, J. (1981). A class of smooth estimators for discrete distributions. *Biometrika, 68*(1), 301–309.

The R Package DAKS: Basic Functions and Complex Algorithms in Knowledge Space Theory

Anatol Sargin and Ali Ünlü

Abstract R is a language and environment for statistical computing and graphics. It is available as free software and can easily be extended by user contributed packages. This paper introduces the R package DAKS (version 1.0-0) developed by the authors for the theory of knowledge spaces in psychometrics. Knowledge space theory provides a theoretical framework for the modeling, assessment, and training of knowledge. It utilizes the idea that some pieces of knowledge may imply others, and is based on order and set theory. The package DAKS contains general functions, for instance, for switching between different formulations in knowledge space theory, a simulation tool, a graph drawing device for Hasse diagrams, and several data analysis methods for detecting implications between test items.

1 Introduction

Knowledge space theory (KST) was introduced by Doignon and Falmagne (1985) and most of the theory is presented in a monograph by Doignon and Falmagne (1999). KST provides a theoretical framework for the modeling, assessment, and training of knowledge. It utilizes the idea that some pieces of knowledge may imply others. Implications between pieces of knowledge are modeled in KST by order and set theoretic structures. Deriving implications from data plays an important role in KST. Three inductive item tree analysis (IITA) algorithms have been proposed for deriving implications from dichotomous data: the original IITA algorithm (Schrepp 2003), and the corrected and minimized corrected IITA algorithms (Sargin and Ünlü 2009; Ünlü and Sargin 2010).

These exploratory data analysis methods constitute the main part of the R (http://www.R-project.org/; R Development Core Team 2010) package DAKS (version 1.0-0), which is available on CRAN (http://CRAN.R-project.org/package=DAKS). This package also implements functions for computing population and estimated

A. Sargin (✉)
Fakultät Statistik, Technische Universität Dortmund, D-44221 Dortmund, Germany
e-mail: sargin@statistik.tu-dortmund.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_28, © Springer-Verlag Berlin Heidelberg 2010

asymptotic variances of the fit measures. Other features are a Hasse diagram drawing device, a simulation tool for generating binary response data and order structures, and functions for switching between item and person related formulations in KST and for computing response pattern and knowledge state frequencies in the data.

2 Basics of KST and IITA

Assume a set Q of m dichotomous items. Mastering an item $j \in Q$ may imply mastering another item $i \in Q$. If no response errors are made, these implications, $j \to i$, entail that only certain response patterns (represented by subsets of Q) are possible. Those response patterns are called knowledge states, and the set of all knowledge states (including $\emptyset$ and Q) is called a knowledge structure, and denoted by $\mathcal{K}$. The knowledge structure $\mathcal{K}$ is a subset of 2^Q, the power set of Q. Implications are assumed to form a quasi order, that is, a reflexive, transitive binary relation, $\sqsubseteq$ on the item set Q. In other words, an implication $j \to i$ stands for the pair $(i, j) \in \sqsubseteq$, also denoted by $i \sqsubseteq j$. Quasi orders are referred to as surmise relations in KST.

Let n be the sample size. The data are the observed absolute counts of response patterns $R \subset Q$. Let D denote the corresponding $n \times m$ data matrix of 0/1 item scores. The data are assumed to be multinomially distributed over 2^Q. Let $\rho(R)$ denote the (unknown) true probability of occurrence of a response pattern R. The basic local independence model (BLIM) is based on the following assumptions. To each knowledge state $K \in \mathcal{K}$ is attached a probability $p(K)$ measuring the likelihood that a respondent is in state K. For a manifest response pattern $R \subset Q$ and a latent knowledge state $K \in \mathcal{K}$, $r(R, K)$ specifies the conditional probability of response pattern R for a respondent in state K. The item responses of a respondent are assumed to be independent given the knowledge state of the respondent (local independence). The response error, that is, careless error and lucky guess, probabilities β_q and η_q are attached to the items and do not vary with the knowledge states.

The BLIM allows expressing the occurrence probabilities $\rho(R)$ of response patterns R by means of the model parameters $p(K)$ and β_q, η_q:

$$\rho(R) = \sum_{K \in \mathcal{K}} \left\{ \prod_{q \in K \setminus R} \beta_q \cdot \prod_{q \in K \cap R} (1 - \beta_q) \cdot \prod_{q \in R \setminus K} \eta_q \cdot \prod_{q \in Q \setminus (R \cup K)} (1 - \eta_q) \right\} p(K).$$

The number of independent model parameters is $2|Q| + (|\mathcal{K}| - 1)$, where $|\mathcal{K}|$ denotes the size of $\mathcal{K}$. Since $|\mathcal{K}|$ generally tends to be prohibitively large in practice, parameter estimation and model testing based on classical maximum likelihood methodology are not feasible in general. This is why exploratory methods such as the IITA algorithms are important in KST.

Random errors in the responses of a respondent are the reason why deriving a surmise relation from data is difficult. The three IITA algorithms are exploratory

methods for extracting surmise relations from data. In each algorithm, competing quasi orders are generated, and a fit measure is computed for every relation in order to find the quasi order that fits the data best. In the following, the IITA algorithms are briefly reviewed (for details, see Sargin and Ünlü 2009; Schrepp 2003; Ünlü and Sargin 2010).

The inductive procedure for generating the competing quasi orders is the same for all three IITA algorithms: For two items i, j, the value $b_{ij} := |\{R \in D | i \notin R \wedge j \in R\}|$ is the number of counterexamples, that is, the number of observed response patterns R in the data matrix D contradicting $j \to i$. Based on these values, binary relations $\sqsubseteq_L$ for $L = 0, \ldots, n$ are defined. Let $i \sqsubseteq_0 j :\Leftrightarrow b_{ij} = 0$. The relation $\sqsubseteq_0$ is a quasi order. Construct inductively: Assume $\sqsubseteq_L$ is transitive. Define $S^{(0)}_{L+1} := \{(i, j) | b_{ij} \leq L + 1 \wedge i \not\sqsubseteq_L j\}$. From $S^{(0)}_{L+1}$, exclude those item pairs that cause an intransitivity in $\sqsubseteq_L \cup S^{(0)}_{L+1}$; the remaining pairs are referred to as $S^{(1)}_{L+1}$. This process continues iteratively, k times, until no intransitivity is caused. The generated relation $\sqsubseteq_{L+1} := \sqsubseteq_L \cup S^{(k)}_{L+1}$ is a quasi order by construction.

The *diff* fit measure is defined by

$$diff(\sqsubseteq, D) = \frac{1}{m(m-1)} \sum_{i \neq j} (b_{ij} - b^*_{ij})^2,$$

where (depending on the algorithm used) corresponding estimates b^*_{ij} are used. These estimates are computed based on a single error probability.

In the original IITA version this single error rate is computed by

$$\gamma_{\sqsubseteq} = \frac{1}{|\sqsubseteq| - m} \sum_{i \sqsubseteq j, i \neq j} \frac{b_{ij}}{p_j n}.$$

If $(i, j) \in \sqsubseteq$, the expected number of counterexamples is estimated by $b^*_{ij} = \gamma_{\sqsubseteq} p_j n$. If $(i, j) \notin \sqsubseteq$, the estimate $b^*_{ij} = (1 - p_i) p_j n (1 - \gamma_{\sqsubseteq})$ is used.

In the corrected IITA version the same $\gamma_{\sqsubseteq}$ and $b^*_{ij} = \gamma_{\sqsubseteq} p_j n$ for $(i, j) \in \sqsubseteq$ are used. The choice for b^*_{ij} in the case of $(i, j) \notin \sqsubseteq$ now depends on whether $(j, i) \notin \sqsubseteq$ or $(j, i) \in \sqsubseteq$. If $(i, j) \notin \sqsubseteq$ and $(j, i) \notin \sqsubseteq$, set $b^*_{ij} = (1 - p_i) p_j n$ (independence between i and j). If $(i, j) \notin \sqsubseteq$ and $(j, i) \in \sqsubseteq$, set $b^*_{ij} = (p_j - p_i + \gamma_{\sqsubseteq} p_i) n$ (follows from derivations in the two-by-two table for i and j).

In the minimized corrected IITA version the corrected estimators b^*_{ij} of the corrected IITA version are used. Minimizing the *diff* expression as a function of the error probability $\gamma_{\sqsubseteq}$ gives $\gamma_{\sqsubseteq} = -\frac{x_1 + x_2}{x_3 + x_4}$, where

$$x_1 = \sum_{i \not\sqsubseteq j \wedge j \sqsubseteq i} -2 b_{ij} p_i n + 2 p_i p_j n^2 - 2 p_i^2 n^2,$$

$$x_2 = \sum_{i \sqsubseteq j} -2 b_{ij} p_j n,$$

$$x_3 = \sum_{i \not\sqsubseteq j \wedge j \sqsubseteq i} 2p_i^2 n^2,$$
$$x_4 = \sum_{i \sqsubseteq j} 2p_j^2 n^2.$$

The idea here is to use the corrected estimators and to optimize the fit criterion. The fit measure then favors quasi orders that lead to smallest minimum discrepancies, or equivalently, largest maximum matches, between the observed and expected numbers of counterexamples.

In Sargin and Ünlü (2009) and Ünlü and Sargin (2010), it is shown that the corrected and minimized corrected estimation schemes lead to better results.

3 The R Package DAKS

In this section, we describe how surmise relations and knowledge structures are implemented, and illustrate some functions of this package with real and simulated data. At the end of the section we give a table summarizing all functions of the package. The implementation in `R` is based on the packages `sets` and `relations`, developed by David Meyer and Kurt Hornik.

A quasi order is a set of tuples, where each tuple is a pair (i, j) representing the implication $j \to i$. The following `R` output shows an example quasi order:

```
{(1, 2), (1, 3), (2, 3)}
```

or

```
{(1L, 2L), (1L, 3L), (2L, 3L)}
```

This code is to be read: item 1 is implied by items 2 and 3, and item 2 is implied by item 3. This gives the chain $3 \to 2 \to 1$. Note that in the second code line an item i is represented by iL. This transformation takes place internally in the packages `sets` or `relations`, but it does not influence computed results. Note that reflexive pairs are not shown in order to reveal implications between different items only, and to save computing time.

We exemplify usage of the package `DAKS` with part of the 2003 Programme for International Student Assessment (PISA; http://www.pisa.oecd.org/) data. The dataset consists of the (1 for correct, 0 for incorrect) answers by 340 German students on a five-item dichotomously scored mathematical literacy test. This is the `pisa` dataset accompanying the package `DAKS`.

First, we get a general idea of the data by looking at the five highest frequencies of occurring response patterns and the numbers of correct answers for all test items.

```
R> pat <- pattern(pisa)
R> pat
$response.patterns
```

```
11100 11000 10000 11110 00000
   67    61    41    40    20

$states
NULL

R> sum(pat$response.patterns)
[1] 229
```

We see that the five most frequent response patterns make up for 229 out of the 340 patterns. These are the Guttman patterns of the chain $d \to c \to b \to a$. This is also indicated by the following code.

```
R> apply(pisa, 2, table)
    a   b   c   d   e
0  51  91 167 261 293
1 289 249 173  79  47
```

From items a to e, the sample item popularities (proportions-correct) are well-differentiated and strictly decreasing. For instance, item a is most popular (most frequently solved), item e is least popular (least frequently solved).

Since we do not know whether the underlying quasi order may or may not be a chain, we run the minimized corrected IITA algorithm on the PISA data.

```
R> mini <- iita(pisa, v = 1)
R> mini
$diff
 [1]  143.53305  137.39922  132.13348  115.37663  120.16808
 [6]  110.48656  82.54234   38.97623   27.56613  107.39041
 [11] 242.37313 1079.05432  2887.72089

$implications
{(1L, 2L), (1L, 3L), (1L, 4L), (1L, 5L), (2L, 3L),
 (2L, 4L), (2L, 5L), (3L, 4L), (3L, 5L)}

$selection.set.index
[1] 9
```

The quasi order with tenth index in the selection set (the selection is not shown here and can be obtained using the DAKS function ind_gen) is a chain. The neighboring quasi orders, including the solution quasi order with index nine produced here (see also Fig. 1), are very close to a chain, and therefore we expect the underlying (true) quasi order to be one of these.

Graphics are convenient to use and they can present information effectively. Hasse diagrams are used in KST for presenting information. A Hasse diagram can be plotted by:

```
R> hasse(mini$implications, 5)
list()
```

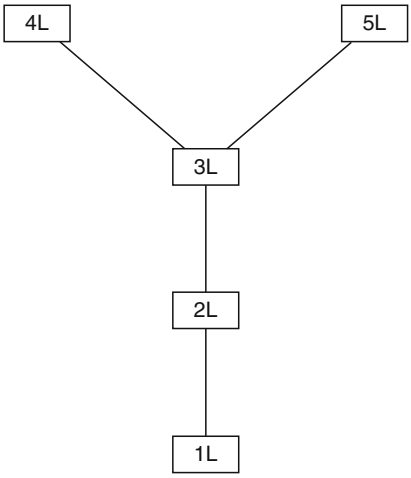


Fig. 1 Hasse diagram of the quasi order obtained for the PISA dataset under the minimized corrected IITA algorithm

To illustrate other functions of the package DAKS, especially functions providing the basis for inferential statistics, we start with simulating a dataset.

```
R> ex_data <- simu(9, 1500, 0.1, 0.1, delta = 0.15)
```

The randomly generated quasi order underlying the simulated data is:

```
R> ex_data$implications
{(1L, 6L), (3L, 1L), (3L, 2L), (3L, 6L), (3L, 7L), (5L, 1L),
 (5L, 2L), (5L, 3L), (5L, 6L), (5L, 7L), (7L, 2L), (7L, 6L),
 (9L, 2L), (9L, 6L), (9L, 7L)}
```

We run the corrected IITA procedure on the simulated dataset (under the other two algorithms the analyses are analogous).

```
R> ex_corr <- iita(ex_data$dataset, v = 2)
```

The quasi order obtained by data analysis is the true quasi order underlying the data.

```
R> ex_corr$implications == ex_data$implications
[1] TRUE
```

Next we discuss the functions which provide the basis for statistical inference methodology. The corrected IITA algorithm can be performed in population quantities, yielding information about the population *diff* values, population occurrence probabilities of response patterns, population error rates, and the inductively generated selection set, by:

```
R> pop <- pop_iita(ex_data$implications, 0.1, 0.1, 9,
R+ dataset = ex_data$dataset, v = 2)
R> attributes(pop)
$names
[1] "pop.diff"  "pop.matrix"  "error.pop"  "selection.set"
```

As shown in Ünlü and Sargin (2010), the MLEs *diff* are asymptotically normal. Large sample normality with associated standard errors can be used to construct confidence intervals for the population values of and to test hypotheses about the *diff* coefficients. For instance, one could test whether one of two quasi orders has a significantly smaller *diff* value in the population. The quasi orders could, for example, be derived from querying experts. In order to do such a test, the asymptotic variances need to be estimated. Population asymptotic variances and consistent estimators of the latter can be computed using the delta method.

The estimated asymptotic variance can be computed by:

```
R> variance(ex_data$dataset, ex_data$implications, v = 2)
 [1] 5.866841e-06
```

The corresponding population asymptotic variance is:

```
R> pop_variance(pop$pop.matrix, pop$selection.set
R+ [[which(min(pop$pop.diff) == pop$pop.diff)]], pop$error.pop
R+ [which(min(pop$pop.diff) == pop$pop.diff)], v = 2)
 [1] 4.176084e-06
```

The sample and population values are quite similar. The sample variance is a consistent estimator for the population variance (convergence in probability).

Table 1 summarizes all functions of the package DAKS.

Table 1 Summary of the DAKS functions

Function	Short description
corr_iita	Computing *diff* values for the corrected IITA algorithm
hasse	Plotting a Hasse diagram
iita	Computing sample *diff* values and the best fitting quasi order for one of the three IITA algorithms selectively
imp2state	Transforming from implications to knowledge states
ind_gen	Inductively generating a selection set
mini_iita	Computing *diff* values for the minimized corrected IITA algorithm
ob_counter	Computing numbers of observed counterexamples
orig_iita	Computing *diff* values for the original IITA algorithm
pattern	Computing frequencies of response patterns and knowledge states
pop_iita	Computing population *diff* values and the selection set for one of the three IITA algorithms selectively
pop_variance	Computing population asymptotic variances
simu	Data simulation tool
state2imp	Transforming from knowledge states to implications
variance	Computing estimated asymptotic variances

4 Conclusion

This paper has introduced the R package DAKS. This package contains several basic functions for KST, and it primarily implements the IITA methods for data analysis in KST. Functions for computing various population values and for estimating asymptotic variances are also contained. These tools provide the basis for statistical inference methodology and for further analyses in KST. We have described the functions of the package DAKS and demonstrated their usage by real and simulated data examples.

In future research, we plan to implement other fit measures, and functions for computing confidence intervals and for performing hypothesis tests for the *diff* (and other) fit measures.

By contributing the package DAKS we hope to have established a basis for computational work in the so far combinatorial theory of knowledge spaces using the R language and environment.

References

Doignon, J.-P., & Falmagne, J.-C. (1985). Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies, 23*, 175–196.

Doignon, J.-P., & Falmagne, J.-C. (1999). *Knowledge spaces*. Berlin: Springer-Verlag.

R Development Core Team. (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Sargin, A., & Ünlü, A. (2009). Inductive item tree analysis: Corrections, improvements, and comparisons. *Mathematical Social Sciences, 58*, 376–392.

Schrepp, M. (2003). A method for the analysis of hierarchical dependencies between items of a questionnaire. *Methods of Psychological Research Online, 19*, 43–79.

Ünlü, A., & Sargin, A. (2010). Maximum likelihood methodology for diff fit measures for quasi orders. *Manuscript submitted for publication*.

Methods for the Analysis of Skew-Symmetry in Asymmetric Multidimensional Scaling

Giuseppe Bove

Abstract The decomposition of any square matrix in symmetric and skew-symmetric components has inspired many methods in asymmetric multidimensional scaling (reviews are provided e.g. in Zielman and Heiser 1996; Bove and Rocci 1999; Borg and Groenen 2005; Saito and Yadohisa 2005). Separate analyzes of the two components seem preferable when symmetry is much more relevant in the data or when we want to represent separately skew-symmetric residuals of statistical models (e.g. symmetry or quasi-symmetry). In this paper two and three-way methods for the analysis of skew-symmetry are reviewed focalizing on multidimensional models with graphical capabilities.

1 Introduction

A skew-symmetric data matrix $\mathbf{N} = [n_{ij}]$ is a square matrix in which:

$$n_{ij} = -n_{ji} \quad , \quad i, j = 1, 2, \ldots, n \tag{1}$$

Matrix $\mathbf{N}$ can be directly observed as debits/credits balance data, preferences, results of matches in a tournament, etc. or it can be derived as the skew-symmetric component of an asymmetric data matrix $\mathbf{\Omega}$ representing flow data, import/export data, confusion rates, etc. In the second case $\mathbf{\Omega}$ is split into a symmetric part $\mathbf{M}$ and a skew-symmetric part $\mathbf{N}$ such that $\mathbf{\Omega} = \mathbf{M} + \mathbf{N}$, where $\mathbf{M} = 0.5(\mathbf{\Omega} + \mathbf{\Omega}')$ and $\mathbf{N} = 0.5(\mathbf{\Omega} - \mathbf{\Omega}')$, and $\mathbf{\Omega}'$ denotes the transpose of $\mathbf{\Omega}$. The sum of squares is similarly decomposed

$$\|\mathbf{\Omega}\|^2 = \|\mathbf{M}\|^2 + \|\mathbf{N}\|^2$$

G. Bove
Dipartimento di Scienze dell'Educazione, Università degli Studi *Roma Tre*, Italy
e-mail: bove@uniroma3.it

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_29,

or in scalar notation

$$\sum_i \sum_j \omega_{ij}^2 = \sum_i \sum_j m_{ij}^2 + \sum_i \sum_j n_{ij}^2$$

because the sum of cross-products vanishes. The previous splitting property holds also for more general types of inner products which were completely characterized by Critchley (1988). Therefore the skew-symmetric part **N** may be viewed and analyzed independently of the symmetric part **M**. Separate analyzes seem preferable, for instance, when methods for the joint representation of the two components fail to represent adequately matrix **N**, because symmetry is much more relevant in the data, or when we want to represent separately skew-symmetric residuals of statistical models (e.g. symmetry or quasi-symmetry). In the following sections we review some methods for multidimensional representation of **N** focalizing on their graphical capabilities.

2 Scalar Product-like Models (Two-Way Case)

Most of the methods proposed for the analysis of two-way skew-symmetric data are particular cases of the following general formulation:

$$n_{ij} = \mathbf{x}'_i \, \mathbf{R} \, \mathbf{x}_j + e_{ij} \tag{2}$$

in which $\mathbf{x}_i$, $\mathbf{x}_j$ are vectors of loadings (or coordinates) respectively for objects i and j, **R** is a square matrix representing underlying relationships of asymmetry and e_{ij} is an error term. Probably the first multidimensional method for representing skew-symmetry was proposed in a pioneering paper by Gower (1977). It is based on the particular form taken by the singular value decomposition (SVD) of **N** (for a proof see e.g. Gower and Zielman, 1992),

$$\mathbf{N} = \mathbf{P}\Delta\mathbf{Q}' = \mathbf{P}\Delta\mathbf{J}\mathbf{P}' \tag{3}$$

where the columns of **P** are the singular vectors, $\Delta = diag(\delta_1, \delta_1, \delta_2, \delta_2, \ldots)$ contains the singular values, with the last diagonal element equal to zero when n is odd, and **J** is a block diagonal matrix with 2×2 matrices

$$\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

along the diagonal and, if n is odd, the last diagonal element conventionally set to one to ensure that **J** is orthogonal. The singular values come in pairs, and for the presence of **J**, if **a**,**b** are left-hand singular vectors associated with the same singular value then **b**,-**a** are the corresponding right-hand singular vectors. In the

two-dimensional case the method proposed by Gower is based on the best *rank-2* approximation of **N** given by the truncated SVD

$$\mathbf{N} = \mathbf{P}_{(2)}\mathbf{\Delta}_{(2)}\mathbf{J}_{(2)}\mathbf{P'}_{(2)} + \mathbf{E} =$$

$$= [\mathbf{p}_1\ \mathbf{p}_2]\begin{bmatrix}\delta_1 & 0\\ 0 & \delta_1\end{bmatrix}\begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}\begin{bmatrix}\mathbf{p}'_1\\ \mathbf{p}'_2\end{bmatrix} + \mathbf{E} = \delta_1(\mathbf{p}_1\mathbf{p}'_2 - \mathbf{p}_2\mathbf{p}'_1) + \mathbf{E} \tag{4}$$

that is a particular case of (2) with $\mathbf{x}_i = [p_{1i}, p_{2i}]'$ and

$$\mathbf{R} = \begin{bmatrix}0 & \delta_1\\ -\delta_1 & 0\end{bmatrix}$$

so that, in scalar notation, we have

$$n_{ij} = \delta_1(p_{1i}\,p_{2j} - p_{1j}\,p_{2i}) + e_{ij} \tag{5}$$

It follows from the last equation that the method represents in a plane the objects with coordinates provided by vectors $\mathbf{p}_1, \mathbf{p}_2$, and the appropriate interpretation of the diagram is not in terms of distances but in terms of areas, in particular it is the area of the triangle that points i and j form with the origin that is proportional to the size of skew-symmetry, whose sign is given by the plane orientation (positive counterclockwise). This solution is arbitrary to the extent that unit vectors $\mathbf{p}_1, \mathbf{p}_2$ may be replaced by any other linear transformation $[\alpha\mathbf{p}_1 + \beta\mathbf{p}_2,\ \gamma\mathbf{p}_1 + \mu\mathbf{p}_2]$ with $\alpha\mu - \beta\gamma = 1$. However this indeterminacy does not affect the area interpretation of the diagram. For the purpose of illustrating interpretation of triangle areas, in Fig. 1 is provided the diagram obtained applying the Gower method to the skew-symmetric component of asymmetric data concerning the e-message traffic among seven specialists in the study of social networks during a period of 18 months (Freeman 1997, Table 3, p. 11). Labels represent the first three letters of specialist names. Positive skew-symmetry characterize the communication flows from the two specialists named Wel and Ber, this means that they sent more messages than they received, especially to the colleague named Fre (large triangle areas). Small skew-symmetry concern communication between the specialists named Dor, Alb and Mul.

Rotational indeterminacy becomes important for the Gower method when we fit more than two dimensions (a *bimension* or a *hedron*). We can restrict to the case of an even number of dimensions, given the particular form taken by the SVD of **N**. In the case of four dimensions (two bimensions) the model takes the form

$$n_{ij} = \delta_1(p_{1i}\,p_{2j} - p_{1j}\,p_{2i}) + \delta_2(p_{3i}\,p_{4j} - p_{3j}\,p_{4i}) + e^*_{ij} \tag{6}$$

and it follows that to deduce the skew-symmetry between objects i and j we have to sum algebraically (weights given by singular values) twice the areas of triangles in two diagrams. This feature discourages the use of more bimensions in the applications.

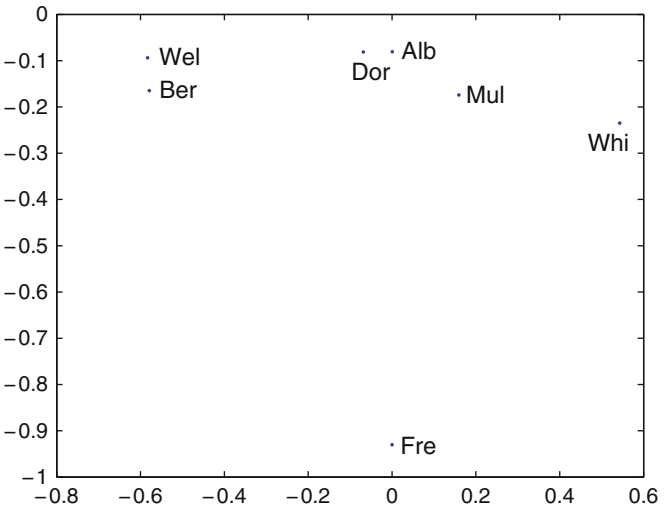


Fig. 1 First bimension of e-mail communication data Freeman 1997

However if we write

$$\mathbf{P}^*_{(r)} = \mathbf{P}_{(r)}\Delta^{1/2}_{(r)} \ , \ \mathbf{U}_{(r)} = \left[\mathrm{p}^*_1, \mathrm{p}^*_3, \ldots\right] \ , \ \mathbf{V}_{(r)} = \left[\mathrm{p}^*_2, \mathrm{p}^*_4, \ldots\right] \tag{7}$$

then the model can be rewritten in matrix form

$$\mathbf{N} = \left[\mathbf{U}_{(r)} \ , \mathbf{V}_{(r)}\right] \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U}'_{(r)} \\ \mathbf{V}'_{(r)} \end{bmatrix} + \mathbf{E} \tag{8}$$

and it turns out (Rocci and Bove, 2002) that rotations *within and between bimensions* are admissible when performed by an orthonormal matrix $\mathbf{T}$ with the particular block form

$$\mathbf{T} = \begin{bmatrix} \mathbf{Z} & \mathbf{W} \\ -\mathbf{W} & \mathbf{Z} \end{bmatrix} \tag{9}$$

that is

$$\begin{aligned} \mathbf{N} &= \left[\mathbf{U}_{(r)} \ , \mathbf{V}_{(r)}\right] \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U}'_{(r)} \\ \mathbf{V}'_{(r)} \end{bmatrix} + \mathbf{E} \ = \\ &= \left[\mathbf{U}_{(r)} \ , \mathbf{V}_{(r)}\right] \mathbf{T} \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{0} \end{bmatrix} \mathbf{T}' \begin{bmatrix} \mathbf{U}'_{(r)} \\ \mathbf{V}'_{(r)} \end{bmatrix} + \mathbf{E} \end{aligned} \tag{10}$$

Rotation methods, like that proposed by Rocci and Bove (2002, p. 415–416), can allow to isolate independent systems of relationships in each bimension (as rotations to simple structure in the factor analytic tradition), reducing the problem to make linear combinations of areas in the different bimensions to reconstruct skew-symmetry (for an application to import/export data see Rocci and Bove 2002).

Rotations where particularly studied in a famous unpublished report by Harshmann (1981) where he proposed Dedicom analysis of matrix **N**. Originally proposed for square asymmetric data matrices, the model takes the form

$$\mathbf{N} = \mathbf{ARA}' + \mathbf{E} \tag{11}$$

The truncated SVD of **N** provides *r-dimensional* Dedicom representation as well ($\mathbf{A} = \mathbf{P}_{(r)}$, $\mathbf{R} = \Delta_{(r)}\mathbf{J}_{(r)}$). Harshmann (1981) proposed several procedures for using rotations to obtain simple structure.

Alternating least squares algorithms were proposed by Kiers and Takane (1993) to fit Dedicom with different constraints on different bimensions. Instead of performing a simple structure rotation of bimensions they proposed to constrain the solution by imposing the same constraints on both dimensions of a bimension. These algorithms can be usefully combined with rotations and *a priori* information on asymmetry in order to define a possible strategy. In fact, when a single bimension is not sufficient to represent adequately the data, we can fit more bimensions, choosing constraints suggested by a preliminary rotated solution.

3 Scalar Product-like Models (Three-Way Case)

Methods for skew-symmetry were also proposed when we have to analyze more data matrices regarding the same set of objects (*three-way* case). Replications may represent several times, different subjects or occasions of observation. A general formulation for the different methods proposed is a natural extension of the two-way case

$$n_{ijk} = \mathbf{x}'_i \, \mathbf{R}_k \, \mathbf{x}_j + e_{ijk} \tag{12}$$

in which $\mathbf{x}_i$, $\mathbf{x}_j$ are vectors of loadings (or coordinates) respectively for objects i and j in a *common space*, $\mathbf{R}_k$ is a square matrix representing underlying relationships of asymmetry at occasion k and e_{ijk} is an error term.

In the following we list and comment briefly some particular cases of model (12), in the equations below $\mathbf{D}_k$ are diagonal matrices with diagonal entries coming in pairs and indicating the relative importance of the underlying bimensions for occasion k.

(a) Three-way Dedicom, skew-Idioscal ($\mathbf{R}_k$ *skew*) (Harshmann, 1978; Kiers, 1993; Zielman, 1993)
(b) Three-way single domain ("strong") Dedicom ($\mathbf{R}_k = \mathbf{D}_k\mathbf{R}\mathbf{D}_k$ *skew*) (Harshmann, 1978; Kiers, 1993)
(c) Skew-Indscal ($\mathbf{R}_k = \mathbf{D}_k\mathbf{J}$) (Zielman, 1993)

Models (a)–(c) were originally proposed for three-way (a)symmetric multidimensional scaling and adapted to the skew-symmetric case. For only one bimension the models are identical. Case (a) is most general but has a difficult interpretation

because matrix $\mathbf{R}_k$ is not necessarily block diagonal, so that when we have more than one bimension we need to analyze all possible pairs of dimensions (i.e. a large number of diagrams). An application and interpretation of model (b) is provided in Lundy et al. (2003). The interpretation of model (c) is very similar to the two-way case, keeping in mind the different scales contained in matrix $\mathbf{D}_k$ corresponding to each occasion. Negative or positive weights can be interpreted in term of direction of asymmetry, unlike symmetric Indscal where negative weights cannot be considered. An application of this method to social mobility data is provided in Zielman (1993).

A three-way model for asymmetric proximities was proposed in Zielman (1991, p. 11), where the skew-symmetric component of the data is represented by a linear model corresponding to the "vector model" of preference data proposed by Carroll (1972). The model has the following form

$$n_{ijk} = \mathbf{w}'_k(\mathbf{x}_i - \mathbf{x}_j) + e_{ijk} \tag{13}$$

where the vector $\mathbf{w}_k$ represents occasion k; the projections of objects-points on the vector indicate the skew-symmetry rank order for that particular occasion.

4 Distance-like Models

The analysis of distances in a diagram is easier than the analysis of areas of triangles. For this reason some authors considered the possibility to model skew-symmetry by distances. Bove (1989) proposed to represent the size of skew-symmetry $\left|n_{ij}\right|$ by euclidean distances performing standard symmetric multidimensional scaling, that is by the model

$$f(\left|n_{ij}\right|) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)} + e_{ij} \tag{14}$$

An advantage of this model is that it is easy to incorporate non metric approaches and external information regarding the objects (Bove 2006) even by standard statistical software (e.g. Proxscal, Spss-Categories). A disadvantage is that we loose the possibility to analyze the sign of skew-symmetry in the diagrams. A proposal to avoid this inconvenience is provided in Borg and Groenen (2005) that represent skew-symmetry by the model

$$n_{ij} = sign(\mathbf{x}'_i\mathbf{J}\mathbf{x}_j)\sqrt{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)} + e_{ij} \tag{15}$$

where distances between points estimate the size of skew-symmetry and the direction of rotation provides the sign, i.e. for each fixed object-point i all points j positioned in the half plane with angles between 0° and 180° (clockwise direction) have a positive estimate for n_{ij}, all the other points positioned in the half plane with angles between 0° and −180° have a negative estimate for n_{ij}. Thus the diagram interpretation of the sign of skew-symmetry works like in *Gower diagrams*. Borg

and Groenen (2005, p. 501–502) also provide an application of the previous model to the largely known Morse-code confusion data to show the performance of their model. However they remind that a general-purpose optimization routine in MatLab was applied to fit model (15), that may be quite sensitive to local optima.

5 Conclusions

We have reviewed some models for the analysis of skew-symmetry in two and three-way cases. A possible strategy of analysis based on rotations and parameter constraining was also suggested in the two-way case when data are represented in more bimensions. Future developments could concern comparative applications of three-way methods and further study of performances of the approaches modeling skew-symmetry by distances.

References

Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling. Theory and applications* (2nd ed.). New York: Springer.

Bove, G. (1989). *New methods of representation of proximity data. Doctoral thesis (in italian)*. Dipartimento di Statistica, Probabilita' e Statistiche Applicate, Universita' La Sapienza, Roma.

Bove, G. (2006). Approaches to asymmetric multidimensional scaling with external information. In S. Zani, A. Cerioli, et al. (Eds.), *Data analysis, classification and the forward search* (pp. 69–76). Berlin: Springer.

Bove, G., & Rocci, R. (1999). Methods for asymmetric three-way scaling. In M. Vichi and O. Opitz (Eds.), *Classification and data analysis. Theory and application* (pp. 131–138). Berlin: Springer.

Carroll, J. D. (1972). Individual differences and multidimensional scaling. In R. N. Shepard et al. (Eds.), *Multidimensional scaling: Theory and applications in the behavioral sciences* (Vol. 1, pp. 105–155). New York: Seminar Press.

Critchley, F. (1988). On characterization of the inner-products of the space of square matrices of given order which render orthogonal the symmetric and the skew-symmetric subspaces. *Warwick Statistics Research Report 171*, University of Warwick, Coventry.

Freeman, L. C. (1997). Uncovering organizational hierarchies. *Computational and Mathematical Organization Theory, 3*, 5–18.

Gower, J. C. (1977). The analysis of asymmetry and orthogonality. In J. R. Barra et al. (Eds.), *Recent developments in statistics* (pp. 109–123). Amsterdam: North Holland.

Gower, J. C., & Zielman, B. (1992). Some remarks on orthogonality in the analysis of asymmetry. *R.R.-92-08*, Department of Data Theory, Leiden.

Harshmann, R. A. (1978). *Models for analysis of asymmetrical relationships among N objects or stimuli*. Paper presented at the First Joint Meeting of the Psychometric Society and the Society for Mathematical Psychology, Hamilton, Ontario.

Harshmann, R. A. (1981). *DEDICOM Analysis of Skew-Symmetric Data. Part I:Theory*. Unpublished technical memorandum, Bell Laboraries, Murray Hill.

Kiers, H. A. L. (1993). An alternating least squares algorithm for PARAFAC2 and DEDICOM3. *Computational Statistics and Data Analysis, 16*, 103–118.

Kiers, H. A. L., & Takane, Y. (1993). Constrained DEDICOM. *Psychometrika, 58*, 339–355.

Lundy, M. E., Harshmann, R. A., Paatero, P., & Swartzman, L. C. (2003). *Application of the 3-way Dedicom model to skew-symmetric data for paired preference ratings of treatments for chronic back pain*. Paper presented at the TRICAP 2003 Meeting, Lexington, Kentucky.

Rocci, R., & Bove, G. (2002). Rotation techniques in asymmetric multidimensional scaling. *Journal of Computational and Graphical Statistics, 11*, 405–419.

Saito, T., & Yadohisa, H. (2005). *Data analysis of asymmetric structures. Advanced approaches in computational statistics*. New York: Marcel Dekker.

Zielman, B. (1991). Three-way scaling of asymmetric proximities. *R.R.-91-01*, Department of Data Theory, Leiden.

Zielman, B. (1993). Two methods for multidimensional analysis of three-way skew-symmetric matrices. *R.R.-93-01*, Department of Data Theory, Leiden.

Zielman, B., & Heiser, W. J. (1996). Models for asymmetric proximities. *British Journal of Mathematical and Statistical Psychology, 49*, 127–146.

Canonical Correspondence Analysis in Social Science Research

Michael Greenacre

Abstract The use of simple and multiple correspondence analysis is well established in social science research for understanding relationships between two or more categorical variables. By contrast, canonical correspondence analysis, which is a correspondence analysis with linear restrictions on the solution, has become one of the most popular multivariate techniques in ecological research. This restricted form of correspondence analysis can be used profitably in social science research as well, as is demonstrated in this paper. We first illustrate the result that canonical correspondence analysis of an indicator matrix, restricted to be related to an external categorical variable, reduces to a simple correspondence analysis of a set of concatenated (or "stacked") tables. Then we show how canonical correspondence analysis can be used to focus on, or partial out, a particular set of response categories in sample survey data. For example, the method can be used to partial out the influence of missing responses, which usually dominate the results of a multiple correspondence analysis.

1 Introduction

Simple correspondence analysis (CA) of two categorical variables, and multiple correspondence analysis (MCA) of more than two variables, are methods commonly used to visualize and interpret categorical data in the social and environmental sciences. In ecology one of the main uses of CA is in a form known as canonical correspondence analysis (CCA), which visualizes a matrix of biological data (e.g., abundance data of various species at a set of sampling locations) in relation to a set of concomitant environmental variables, which could be measured on continuous and/or discrete scales (CCA was originally proposed in Ter Braak 1986; for a short summary, see Greenacre 2007, Chap. 24). In CCA the solution space, usually

M. Greenacre
Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005 Barcelona
e-mail: michael@upf.es

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_30,

a two-dimensional plane, is not the optimal one that would have been obtained by regular CA, but is restricted to be related linearly to the concomitant variables – in other words, the objective is to find a solution directly related to the concomitant variables.

This idea can also be used fruitfully in the analysis of social science data, as we shall demonstrate. We give two possibilities in the context of MCA of a set of question responses in a social survey: first, the analysis of the questions with a single concomitant variable that is discrete; and second, the focusing on, or partialling out, a chosen set of response categories. The strategy of partialling out the effects of missing responses in a questionnaire survey is particularly useful since these usually dominate the MCA solution and obscure the more interesting relationships amongst the substantive variables.

2 Canonical Correspondence Analysis

The theory of CA is well-known and we just summarize it here to establish notation. Suppose that $\mathbf{N}$ is an $I \times J$ table of non-negative data – divided by its grand total n it is called the *correspondence matrix* $\mathbf{P} = (1/n)\mathbf{N}$. Let the row and column marginal totals of $\mathbf{P}$ be the vectors $\mathbf{r}$ and $\mathbf{c}$ respectively – these are the weights, or *masses*, associated with the rows and columns. Let $\mathbf{D}_r$ and $\mathbf{D}_c$ be the diagonal matrices of these masses. Then CA is based on the singular-value decomposition (SVD) of $\mathbf{S} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^\mathsf{T})\mathbf{D}_c^{-1/2}$: $\mathbf{S} = \mathbf{U}\mathbf{D}_\sigma\mathbf{V}^\mathsf{T}$, where $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}$. The *principal coordinates* of the rows and columns are $\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\mathbf{D}_\sigma$ and $\mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{V}\mathbf{D}_\sigma$ respectively, hence are scaled in such a way that $\mathbf{F}^\mathsf{T}\mathbf{D}_r\mathbf{F} = \mathbf{G}^\mathsf{T}\mathbf{D}_c\mathbf{G} = \mathbf{D}_\sigma^2 := \mathbf{D}_\sigma^2$, i.e. the weighted sum of squares of the coordinates on the k-th dimension (or their inertia in the direction of this dimension) is equal to σ_k^2, called the *principal inertia* (or eigenvalue) on dimension k. *Standard coordinates* are similarly defined but without scaling on the right by the singular values, and hence the standard coordinates on any given dimension have weighted sum of squares equal to 1. The sum of squares of the decomposed matrix $\mathbf{S}$ is called the *total inertia*, and this quantity is decomposed by the squared singular values σ_k^2, which are in decreasing order. The best solution in two dimensions would use the first two columns of the coordinate matrices, and the explained inertia would be the sum of the first two terms $\sigma_1^2 + \sigma_2^2$, usually expressed as a percentage of the total inertia.

When a separate set of concomitant variables is available that can be regarded as possibly explaining the phenomena evident in the results of a CA, it is common to relate them to a given CA solution as *supplementary variables* (see, for example, Greenacre 2007: Chap. 12). In ecological applications this is known as 'indirect ordination' because the concomitant variables play no role in determining the solution but are mapped into the solution a posteriori, with the result that the concomitant variables may be poorly correlated with the CA solution. By contrast, in CCA, the dimensions are intentionally defined as linear combinations of the concomitant variables, so this ensures that the concomitant variables have high

correlations with the solution space: this is called 'direct ordination'. Geometrically, the principal axes in CCA are sought in that restricted part of the space which is projected onto the concomitant variables. We can also look for principal axes in the space that is uncorrelated with the concomitant variables, in which case the (linear) effects of the concomitant variables have been partialled out. In this latter case we have what is called partial canonical correspondence analysis (PCCA), which could optionally also involve its own separate set of constraining concomitant variables.

Algebraically, CCA follows the same scheme as CA except that there is an initial projection of the data onto the space spanned by the concomitant variables. Suppose $\mathbf{X}(I \times K)$ is the matrix of K concomitant variables used to restrict the CA solution, supposed to be standardized to mean 0, variance 1 (the rows are always weighted by their masses in all computations). Then the projection matrix is $\mathbf{Q} = \mathbf{D}_r^{1/2}(\mathbf{X}^\mathsf{T}\mathbf{D}_r\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{D}_r^{1/2}$ and the matrix $\mathbf{S}$ defined previously, projected onto the concomitant variables, is $\mathbf{S}^* = \mathbf{QS}$. Notice here that projection, which is a scalar product operation, incorporates the weighting of the rows by the row masses in $\mathbf{D}_r$. Having performed the projection, everything follows as for regular CA, using $\mathbf{S}^*$ rather than $\mathbf{S}$. For PCCA, projection takes place on the space orthogonal to the concomitant variables: $\mathbf{S}^\perp = (\mathbf{I} - \mathbf{S})\mathbf{Q}$, and then the same steps follow as before, applied to $\mathbf{S}^\perp$.

In CCA there is a double decomposition of inertia: first, total inertia is decomposed into a part in the restricted space and the complementary part in the unrestricted space. In the restricted space there is the usual decomposition along principal axes, and similarly there can be a decomposition of the complementary part of inertia along principal axes in the unrestricted space. In the applications considered here, we shall use these results in the case of MCA, when the primary data in $\mathbf{N}$ consist of dummy variables. Hence, to make our terminology even more specific, we could say that we are performing 'canonical multiple correspondence analysis' and 'partial canonical multiple correspondence analysis'. The data considered are from the survey of International Social Survey Program (ISSP) on Family and Changing Gender Roles II (ISSP 1994), specifically responses from 2,494 respondents in Spain to 11 questions relating to the issue of working women (Table 1 lists the questions and the five substantive response categories).

3 Constraining by a Single Categorical Variable

In social science applications, the variables being analyzed are generally categorical, hence the relevance of CA and MCA. Figure 1 shows the MCA of the Spanish data for the questions in Table 1. Three clusters of response categories are evident: all the missing categories at upper right, all the moderate responses ("agree" and "disagree") and middle responses ("neither agree nor disagree") in a bunch near the origin (these are the most frequent responses), and all extreme responses ("strongly agree" and "strongly disagree") at upper left. A demographic variable, age group, with six categories from young to old, a1 (16–25 years) to a6 (more than 65 years),

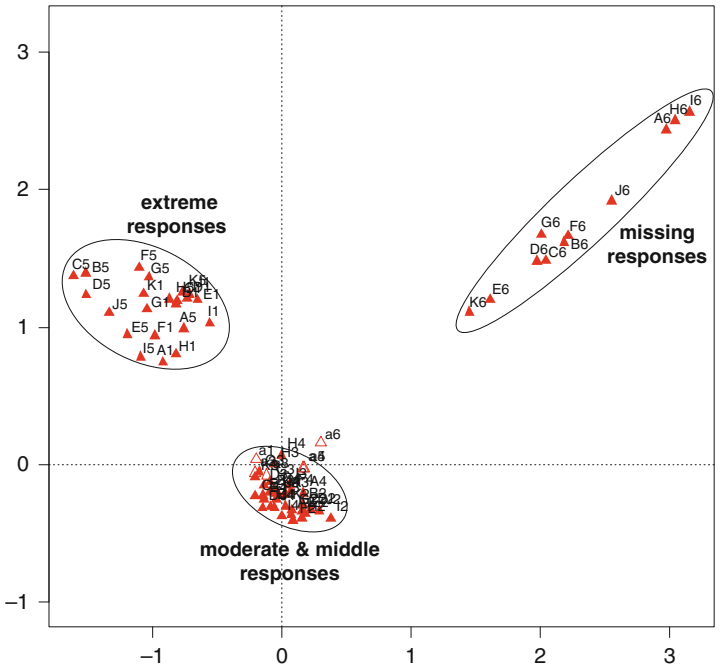


Fig. 1 MCA of 11 questions from ISSP (1994), Spanish sample ($N = 2{,}494$), with age group variable as supplementary – the supplementary age categories are all close together near the centre of the map (e.g., the labels a1, a5 and a6 are just visible, with the oldest age group a6 tending in the direction of the missing responses)

Table 1 Eleven questions from ISSP (1994) concerning women working: respondents had to choose between 1–strongly agree, 2–somewhat agree, 3–neither agree nor disagree, 4–somewhat disagree, 5–strongly disagree

A[+]	A working mother can establish just as warm and secure a relationship with her children as a mother who does not work
B[−]	A pre-school child is likely to suffer if his or her mother works
C[−]	All in all, family life suffers when the woman has a full-time job
D[−]	A job is all right, but what most women really want is a home and children
E[?]	Being a housewife is just as fulfilling as working for pay
F[+]	Having a job is the best way for a woman to be an independent person
G[?]	Most women have to work these days to support their families
H[+]	Both the man and woman should contribute to the household income
I[−]	A man's job is to earn money; a woman's is to look after the home & family
J[?]	It is not good if the man stays at home and cares for the children and the woman goes out to work
K[?]	Family life often suffers because men concentrate too much on their work

Don't know/missings are coded as 6. Statements clearly in favour of women working are marked (+), others clearly opposed (−), and the remainder not so clearly oriented (?)

is displayed in the form of supplementary points, all near the origin. This result is typical of an MCA of questionnaire data such as these: the missing responses dominate as well as response styles (moderates versus extremes, independent of the fact that several questions have reverse wording) and a supplementary variable has categories only slightly separated spatially.

Suppose that we wanted to see the map of the response categories specifically in their relation to the age groups. This can be achieved by constraining the solution space to be defined by the age categories, that is performing a CCA on the indicator matrix of the 11 questions (66 dummy variables), with the indicator matrix of the age groups (six dummy variables) as the constraining variables. This CCA is identical to the CA of the concatenated matrix of the 11 cross-tabulations of the questions with the age variable, that is the matrix with 66 rows and six columns with the 11 cross-tables stacked one on top of another (this stacked matrix is equal to the transpose of the 66-column indicator matrix of the questions multiplied by the six-column indicator matrix of the age groups). This result follows from the fact that CCA is equivalently defined as the CA of the weighted averages of the constraining variables for each response category (see, for example, Greenacre 2007, 191–192). This simplifying result appears to be not well-known: for example, Nishisato's "forced classification" (Nishisato 1984) is identical to the CCA described here, which in turn is identical to the CA of the stacked tables. Figure 2 shows the CA of the stacked

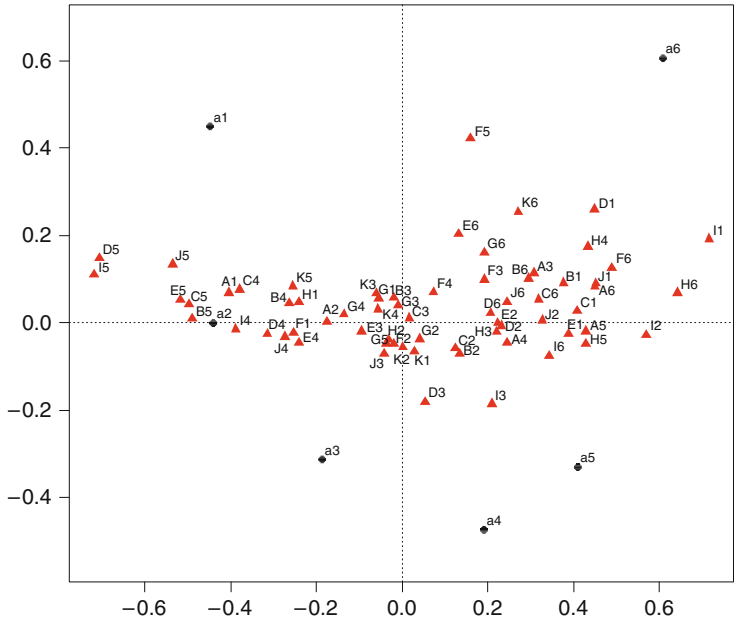


Fig. 2 CA of cross-tabulations of 11 questions with age groups. The standard biplot scaling is used (Greenacre 2007, Chap. 13)

tables, which is more efficiently performed than the CCA of the large indicator matrices.

In Fig. 2 the domination of the response styles seen in Fig. 1 has vanished and we pick up the liberal-to-traditional scale from left to right in the response categories, with the reversely worded questions lining up as we would expect: for example, the most spread-out question is in favour of men working and women staying at home, question I (see Table 1), from I5 on the left to I1 on the right, while question A in favour of women working goes in the opposite direction. Notice that all the missing value categories are on the right, in the direction of the older respondents.

4 Constraints for Dealing with Missing Responses

CCA can be used to focus on, or partial out, an external variable or variables. In Fig. 1 we have all the missing response categories defining a diagonal spread of points, very dominant in the analysis because of the high association amongst missing responses on different questions. To avoid deleting cases that have missing responses from the study, Greenacre and Pardo (2006a,b) proposed a subset version of correspondence analysis to choose subsets of categories for visualizing – this approach can be used to select all substantive response categories and ignore the missing ones. The present approach is an alternative strategy where we define external variables for constraining the solution. There are different ways of doing this, and we show just one of the alternatives where the constraining variable is defined as the count of missing responses for each respondent. For example, a respondent with no missing respondents gets value 0, with one missing response 1, and so on, with respondents giving missing responses to all 11 questions getting a value 11. If we constrain the MCA solution to be linearly related to this single variable we obtain a one-dimensional CCA solution. In Matschinger and Angermeyer (2006) the missing value counts are also used to take care of missing responses – the count variable is added as a categorical variable (i.e., with as many categories as levels of counts) to each of the questions of the questionnaire and then generalized canonical analysis is used with a restriction to concentrate the missing count categories onto a single dimension. The idea is the same: to partial out the missing responses to avoid having to delete cases with missing data.

Figure 3 shows the constrained solution as the horizontal axis (labeled ‘CCA1’), and the second axis is the optimal first axis of the unconstrained solution (labelled ‘CA1’). Comparing this map to Fig. 1 we see that the constraint has forced the missing categories to coincide with the first axis. The variable “missings” that we created, is the sum of the 11 columns of the indicator matrix corresponding to the missing categories, hence its position in space is the average of these categories, as shown by the vector in Fig. 3.

In this sense CCA is acting like a target rotation of the MCA solution. The remaining unconstrained dimensions are orthogonal to this dimension and so the

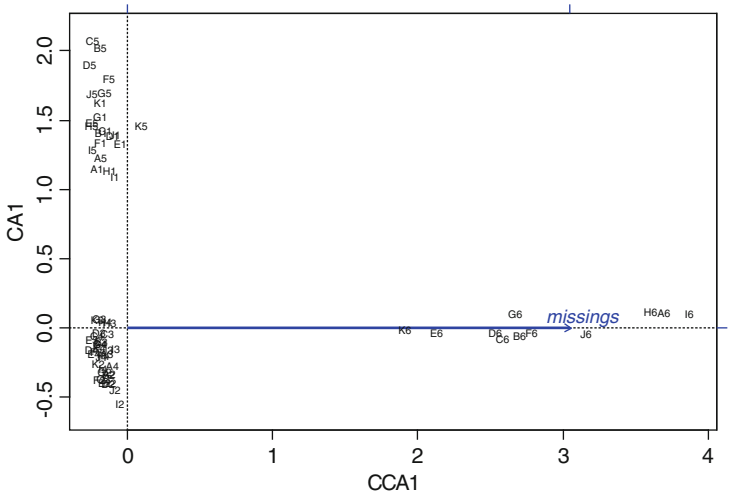


Fig. 3 CCA of 11 questions constrained by number of missings, which is a point vector lying at the average of the 11 dummy variables for the missing categories. This vector is constrained to be the first axis in the CCA

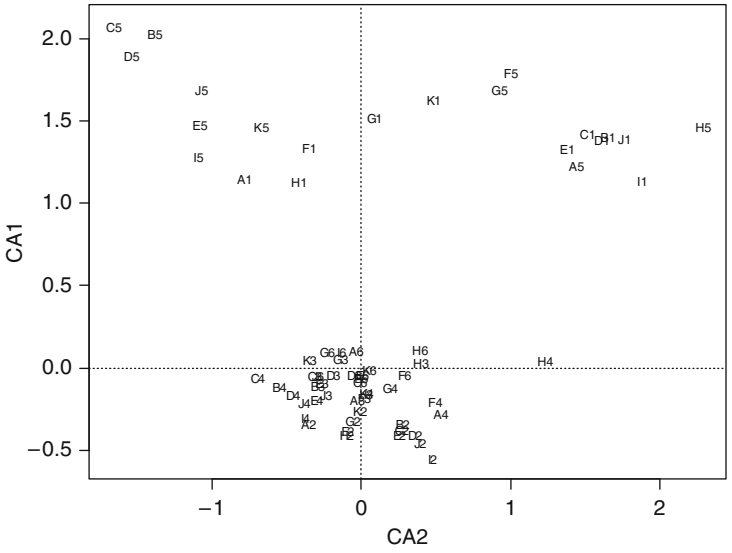


Fig. 4 CCA of 11 questions after partialling out "missings" counts. Missing categories (numbered '6') now play almost no role in the solution

missing response effect has been mostly partialled out. Figure 4 shows axes 2 and 3, the first two unconstrained axes of the CCA – we have kept the first unconstrained axis vertical as in Fig. 3, so that Fig. 4 is a rotation of the solution in Fig. 3 around the vertical axis, bringing into view the next dimension (labeled 'CA2') on the

horizontal axis. The vertical separation is the more important one, separating out the response styles, but now we manage to recover the liberal-traditional dispersion along the horizontal axis, among the extreme responses at the top, and among the moderate and middle responses at the bottom.

5 Discussion

We have shown how CCA can be used to incorporate external information into MCA results or to treat specific response categories in survey data by imposing linear constraints on the solution space. The map can be concentrated on the display of these variables or categories, or their effects can be partialled out. We are also using this approach fruitfully to study the "middle" response categories (Greenacre and Pardo 2008) and their relationship to demographic variables, as well as to partial out acquiescence effects which are rife in questionnaire data.

Acknowledgements This research has been supported by the Fundación BBVA, Madrid, Spain. Partial support of Spanish Ministry of Education and Science grants MTM2008-00642 and MEC-SEJ2006-14098 is also hereby acknowledged.

References

Greenacre, M. (2007). *Correspondence Analysis in Practice* (2nd ed.). London: Chapman & Hall/CRC, 2007. Published in Spanish translation as *La Práctica del Análisis de Correspondencias*, Fundación BBVA, Madrid.

Greenacre, M., & Pardo, R. (2006). Subset correspondence analysis: Visualization of selected response categories in a questionnaire survey. *Sociological Methods and Research, 35*, 193–218.

Greenacre, M., & Pardo, R. (2006). Multiple correspondence analysis of subsets of response categories. In M. Greenacre & J. Blasius (Eds.), *Multiple correspondence analysis and related methods* (pp. 197–217). London: Chapman & Hall/CRC Press.

Greenacre, M., & Pardo, R. (2008). Positioning the "middle" categories in survey research: A multidimensional approach. *Keynote address at the joint conference of the European Methodology Association and the SMABS*, Oviedo, Spain.

ISSP. (1994). Family and changing gender roles II. *International social survey programme*.

Matschinger, H., & Angermeyer, M. C. (2006). The evaluation of "dont know" responses by generalized canonical analysis. In M. Greenacre & J. Blasius (Eds.), *Multiple correspondence analysis and related methods* (pp. 283–298). London: Chapman & Hall/CRC Press.

Nishisato, S. (1984). Forced classification: A simple application of a quantification technique. *Psychometrika, 49*, 25–36.

Ter Braak, C. J. F. (1986). Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. *Ecology, 67*, 1167–1179.

Exploring Data Through Archetypes

Maria Rosaria D'Esposito, Giancarlo Ragozini, and Domenico Vistocco

Abstract In this paper we propose a mixed analytical and graphical exploratory strategy based on data archetypes for the exploratory analysis of multivariate data. Our approach is of considerable help in exploring the periphery of the data scatter, exploiting an outward-inward perspective, to highlight small peripheral groups as well as anomalies, outliers and irregularities in the data cloud shape. The strategy is carried out in a comprehensive quantitative programming environment provided by the joint use of the software system R and of the visualization system GGobi. It provides a visualization system involving both static and dynamic graphics based on the so-called multiple views paradigm. The views are organized in a spreadplot and heavily exploit dynamics and interactive statistical graphics.

1 Introduction

Exploratory data analysis (EDA) is, in the words of Tukey (1977) p. 21, "a detective work, finding and revealing the clues", i.e. uncovering unanticipated structures in the data. EDA uses numerical as well as visual and graphical techniques to accomplish its aims. Graphical and visual tools (such as stars, glyphs, parallel coordinates) become particularly necessary for the exploration of multivariate data as they make it possible to visualize multidimensional data in 2D (for a review see, among others, Chambers et al. 1983; Wegman and Carr 1993).

However, visualization systems that focus on the graphical representation of information can run into several problems. Mainly, there can be loss of valuable information – when too much data are visualized on the screen – and needs to organize discoveries off line (Yang et al. 2007). Interactive and dynamic statistical graphics which incorporate motion and user interaction with graphical display – such as brushing and slicing, coloring and rendering, empirical and algebraic linking

G. Ragozini (✉)
Department of Sociology, Federico II University of Naples, Italy
e-mail: giragoz@unina.it

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_31,

(Young et al. 1993) – can become part of the visualization techniques to enhance the results of the analysis and to overcome some of the problems of the visualization systems. A major advance can also be obtained by introducing integrated analytical devices into the visualization systems, which could aid users in the knowledge discovery task.

In this paper we propose a mixed analytical and graphical exploratory strategy based on data archetypes for the exploratory analysis of multivariate data. Archetypes are few pure types given by weighted average of the data. They are useful in summarizing the data, but, in our opinion, may also be profitably employed to explore the periphery of the data scatter.

It is known that small peripheral groups, anomalies, outliers and irregularities in the data cloud shape in higher dimensions can easily hide in marginal projection or in usual graphical representations. On the contrary, by adopting an outward-inward perspective, i.e. by analyzing the archetypes' surroundings, all the previous data structures can be more easily detected.

With this purpose, we propose an integrated strategy: first, based on the aims of the analysis, extract the archetypes from the data, then represent the data in the space spanned by the archetypes adopting dynamic and interactive visualization tools. The strategy, carried out in the comprehensive quantitative programming environment provided by the joint use of R (R Development Core Team 2008) and GGobi (Cook and Swayne 2007; Lang et al. 2008), will result in a visualization involving both static and dynamic graphics based on the so-called multiple views paradigm (Wilhelm 2005), which allows interaction through dynamic graphics and provides multiple different and simultaneous plots of the same data. The views will be organized in a spreadplot, a spreadsheet-like arrangement of linked, dynamic, interactive plots (Young et al. 1992, 1993).

The paper is organized as follows: Sects. 2 and 3, respectively, present the basic elements of archetypal analysis and spreadplot design; the proposed procedure is illustrated in Sect. 4, both theoretically and practically, exploiting a real dataset; concluding remarks are in Sect. 5.

2 Elements of Archetypal Analysis

Archetypal analysis is a quite recent statistical method for multivariate data analysis (Cutler and Breiman 1994). It aims at finding archetypes that represent a sort of "pure individual types", i.e. few points lying on the boundary of the data scatter that are intended as a synthesis of the observed points. At the same time, as they are not necessarily observed points, they represent ideal objects on which the observed data may be patterned.

Formally, the archetypes $\mathbf{a}'_j$ are a convex combination of the observed data:

$$\mathbf{a}'_j = \boldsymbol{\beta}'_j \mathbf{X} \tag{1}$$

where $\mathbf{X}$ is the observed data matrix, $\boldsymbol{\beta}_{ji} \geq 0 \quad \forall j, i$ and $\boldsymbol{\beta}'_j \mathbf{1} = 1 \quad \forall j$.

On the other hand, all the data points can be expressed in terms of the archetypes:

$$\mathbf{x}_i' = \gamma_i' \mathbf{A} \tag{2}$$

with $\gamma_{ij} \geq 0 \ \ \forall i, j$ and $\gamma_i' \mathbf{1} = 1 \ \ \forall i$. In (2) $\mathbf{A}$ is the archetype matrix and γ_i' are weights of the archetypes for each data point.

Equation (1) and the related constraints on β's coefficients imply that archetypes belong to the convex hull boundary of the data, while (2) and the related constraints on γ's coefficients imply that all the data belong to the convex hull boundary of the archetypes. Hence, the archetypes must coincide with the v vertices of the data convex hull to fulfill the previous conditions (Porzio et al. 2008).

However, in practice, the number of the data convex hull vertices is generally too large to properly synthesize the data. For this reason, looking for a smaller number of pure types, and wishing to preserve their closeness to the data, Cutler and Breiman (1994) defined the archetypes as those m, with $m \leq v$, points that fulfill (2) as far as possible, satisfying all the other conditions. Hence, given m, the archetypes can be defined as the points $\mathbf{A}(m) = (\mathbf{a}_1, \ldots, \mathbf{a}_m)$ minimizing the distances between the observed data points $\mathbf{x}_i'$ and the reconstructed points $\tilde{\mathbf{x}}_i'(m)$, with $\tilde{\mathbf{x}}_i'(m) = \gamma_i'(m) \cdot \mathbf{A}(m)$.

Formally, the archetypes are those points that minimize the quantity:

$$RSS(m) = \left\| \mathbf{X} - \tilde{\mathbf{X}}(m) \right\|_F = \left\| \mathbf{X} - \boldsymbol{\Gamma}(m)\mathbf{A}(m) \right\|_F = \left\| \mathbf{X} - \boldsymbol{\Gamma}(m)\mathbf{B}'(m)\mathbf{X} \right\|_F \tag{3}$$

holding all the other conditions, and where $\|\mathbf{Y}\|_F = \sqrt{Tr\,(\mathbf{YY}')}$ is the Frobenius norm for a generic matrix $\mathbf{Y}$, with $\mathbf{B}(m) = [\beta_{ji}]$ and $\boldsymbol{\Gamma}(m) = [\gamma_{ij}]$. The solution to this minimization equation depends on m, and solutions are not nested as m varies. That is, the archetypal points that solve (3) for $m = m^*$ are not necessarily a subset of the points that solve (3) for $m = m^* + 1$. For this reason, we denote with $\mathbf{a}_j'(m)$ the j-th archetype for a given m, and we will generally have $\mathbf{a}_j'(m) \neq \mathbf{a}_j'(l)$, for $m \neq l$. Theoretically, the $RSS(m)$ in (3) is a decreasing function of m that has the maximum for $m = 1$ and goes to zero for m approaching the number of the convex hull vertices. For a given m, it highlights the synthesizing power of the archetypes since it shows how well archetypes reconstruct data.

Figure 1 exhibits a set of 50 simulated data points with seven convex hull vertices. For this dataset, the $RSS(m)$ function shows that three archetypes are sufficient to synthesize the data. Indeed, for $m = 3$ the $RSS(m)$ is close to zero as the majority of the data belongs to the archetypes' convex hull (the triangle highlighted in Fig. 1), and it is well reconstructed by the archetypes.

Up to now archetypal analysis has been applied in many fields. In the field of physics, it has been used to detect clusters of cellular flames (Stone and Cutler 1996; Stone 2002) and of galaxy spectra (Chan et al. 2003). It has found application as a tool for image decomposition (Marinetti et al. 2006, 2007), where archetypal analysis seems to provide results which are easily interpretable in terms of physical meaning.

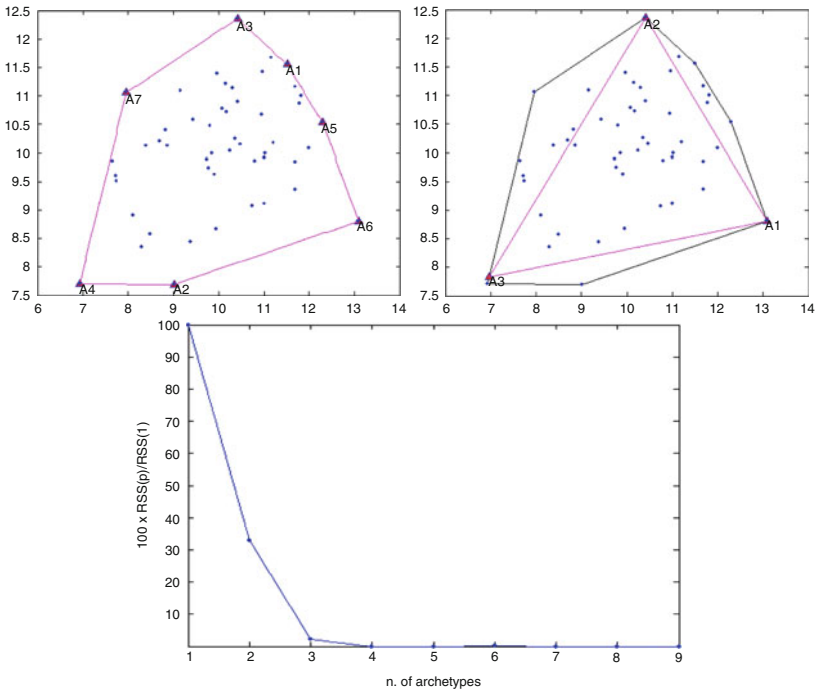


Fig. 1 A set of 50 simulated data points with seven convex hull vertices. *Clockwise*: the dataset with the convex hull boundary and $m = 7$ archetypes highlighted; the dataset with three archetypes highlighted, data convex hull boundary and archetype convex hull boundary; the $RSS(m)$ function

Marketing research is also a relevant field of applications. The idea of archetypes has been associated to the idea of archetypal consumers; they have been exploited for market segmentation and consumer fuzzy clustering (Elder and Pinnel 2003; Li et al. 2003). In the same field, the extension of archetypal analysis to interval coded data has been recently proposed (D'Esposito et al. 2006).

In performance analysis, archetypes have been exploited to construct data driven benchmarks (Porzio et al. 2008), to analyze CPU performance (Heavlin 2007), and to obtain a multivariate ordering procedure based on the idea of the *"worst-best" direction* selected through the archetypes (D'Esposito and Ragozini 2008).

In all the above mentioned applications the archetypes have been mainly adopted as summarizing observations. In this paper we propose to use them as a tool to analyze the structure of the data in a genuine exploratory fashion by merging the analytics of archetypes with dynamic and interactive visualization tools according to a spreadplot scheme (Young et al. 1992).

3 Elements of Spreadplot Design

Enhanced man-machine interaction makes it possible to design tools for an interactive visual exploration of data, and for visually querying the data. Strictly related to interactivity is the paradigm of the linked views (Wilhelm 2005). It consists of linking, empirically or algebraically, several views of the same dataset and in propagating information through different plots that display different aspects (dimensions) of the same dataset (Stuetzle 1987; Young et al. 1993). Brushing, slicing and coloring allow one to visually explore data and to investigate if a particular pattern or position is confirmed on different views of the same dataset.

In this framework, different data views can be arranged as a spreadplot (Young et al. 1992), a spreadsheet-like arrangement of dynamic and interactive plots empirically and algebraically linked together by equations. Each individual view can be either dynamic or static, and can support high-interaction direct manipulation. Moreover, different views can be linked and when the user changes the information shown in one view, the changes can be processed by a set of equations and instantly represented in the others. The user can interactively modify analysis parameter by acting on user interface tools of the spreadplot, such as moving points, sliding cursors.

In this paper, we propose an exploratory strategy based on the archetypes and visually translated into a spreadplot. This includes a set of ad hoc *R* routines implementing the method and is based on open source software whose core is the *R* software system integrated with the visualization system GGobi (Buja et al. 2003) through the *rggobi* package (Lang et al. 2008). We choose the GGobi system since it contains several dynamic and interactive graphics such as tourplot, scatterplot, bar chart and parallel coordinate plot. The whole software architecture exploits the Model-View-Controller (MVC) design pattern and the Observer design pattern (Buschmann et al. 1996; Gamma et al. 1995).

4 The Proposed Exploratory Data Analysis Strategy

It is well known that for high dimensions data structures anomalies and outliers, irregularities in the data cloud shape, and small peripheral groups can easily hide in marginal projections or in usual graphical representations. At the same time, archetypes which are located on the boundary of the data convex hull, can provide an outward-inward point of view on the data scatter that will allow to explore the data cloud peripheries and highlight many data patterns more easily.

The strategy we propose for the exploration of a multivariate dataset is based on the outward-inward perspective given by the archetypes, combined with the geometric properties of the γ coefficients in (2), and the dynamic and interactive visualization tools conveyed in a spreadplot.

The proposed exploratory strategy consists of the following steps:

- Derive the archetypes by increasing their number m one unit at time and look at the $RSS(m)$ function in order to understand the synthesizing power of each additional archetype;
- For each m analyze the archetypes in the data space through some graphical representation (e.g. percentile profile plots, star plot, parallel coordinate plot) to interpret and compare the archetypes;
- On the basis of the previous steps, choose the first interesting m and for this:
  - Represent, through a parallel coordinate plot, the data within the space spanned by the archetypes according to the γ coefficients derived from (2);
  - Use interactive and dynamic tools like brushing, coloring and linking to highlight peripheries selecting the data with γ_{ij} coefficients close to 1 on each archetype, i.e. select the outer data looking for gaps and peripheral structures and for isolated data points;
- Iterate by increasing m, i.e. selecting another subsequent interesting m;
- Stop when increasing m does not provide additional information.

The previous steps are detailed in the following subsections using a real dataset adopted as an illustrative example. The data refer to a study on the performance of central processing units (CPUs) and consist of a set of 209 CPUs to be compared by considering seven performance indicators (Cycle time – ns, Minimum memory – kb, Maximum memory – kb, Cache size – kb, Minimum channels, Maximum channels, Relative performance); low values of the indicators stand for poor performances (Ein-Dor and Feldmesser 1987). Apart from the usual goals (discovering patterns, groups, outliers, ...), the analysis will aim also at finding CPUs with good or bad performance in terms of the seven indicators, and in comparing all the others with them.

4.1 Deriving and Analyzing Archetypes by Varying m

The first step consists of looking at the $RSS(m)$ function defined in (3) and deriving the archetypes as m increases. When the aim is data synthesis, it is necessary to choose one appropriate m. This will corresponds to the one that does not yield a significant decrease in the $RSS(m)$. If the aim is to explore data, we suggest to look at archetypes for different values of m, evaluating also their interpretability by visualizing them through percentile profile plot, stars or other graphical representations. In particular, the percentile profile plot displays for each archetype a sequence of vertical bars (one for each variable) with heights equal to the value of the cumulative empirical distribution evaluated at the archetypal point. It visualizes the archetype's relative standing with respect to the others points.

Figures 2 and 3 show two spreadplots, respectively for $m = 4$ and $m = 6$, portraying the $RSS(m)$ function, percentile profile plots/star plots along with the

parallel coordinate plot for the computer data. Note that in the spreadplots the sliding cursor makes it possible to change interactively the number of archetypes in order to see how percentile profiles and stars change. The values of $m = 4$ and $m = 6$ have been chosen because, by looking at the $RSS(m)$ function in Fig. 3, it appears that $RSS(m)$ decreases sharply up to $m = 6$ with a quite flat behavior between $m = 2$ and $m = 4$.

An inspection of the percentile profile plot in Fig. 2 highlights that archetypes $a_2(4)$ and $a_3(4)$ correspond to bad performances except for the first indicator in $a_3(4)$, as all the percentile bars are low in values, i.e. the archetypes are close to the low values of the indicators. On the other hand $a_4(4)$ represents the best CPUs and $a_1(4)$ the average CPUs. The coordinate parallel plot shows that the majority of CPU data are close together with very few of them having better performances on different indicators. The parallel coordinate plot also highlights skewed marginal distributions.

With $m = 6$ (Fig. 3) we note that one archetype represents the best CPUs ($a_3(6)$), and two archetypes represent the worst CPUs ($a_4(6)$ and $a_5(6)$). The remaining three archetypes represent CPUs with high values on only some indicators. Even if $RSS(6)$ is much lower than $RSS(4)$, the two additional archetypes give no further information and there does not appear to be much gain in interpretability.

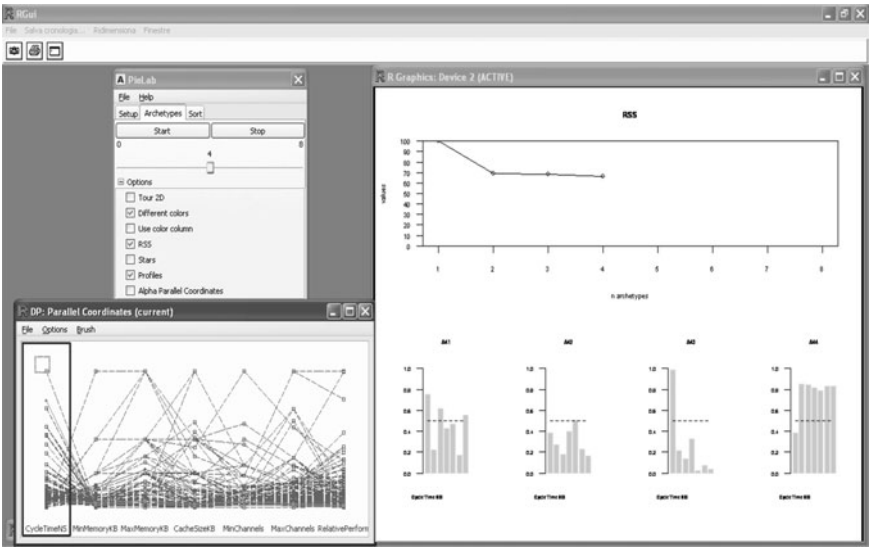


Fig. 2 A spreadplot for the CPUs dataset for $m = 4$ archetypes. *Clockwise*: the control panel which clearly shows the sliding cursor for interactively choosing the number of archetypes, and the list of possible graphical representations; the $RSS(m)$ function and the percentile profile plots for the chosen four archetypes; the parallel coordinate plot of the CPUs data along with the archetypes

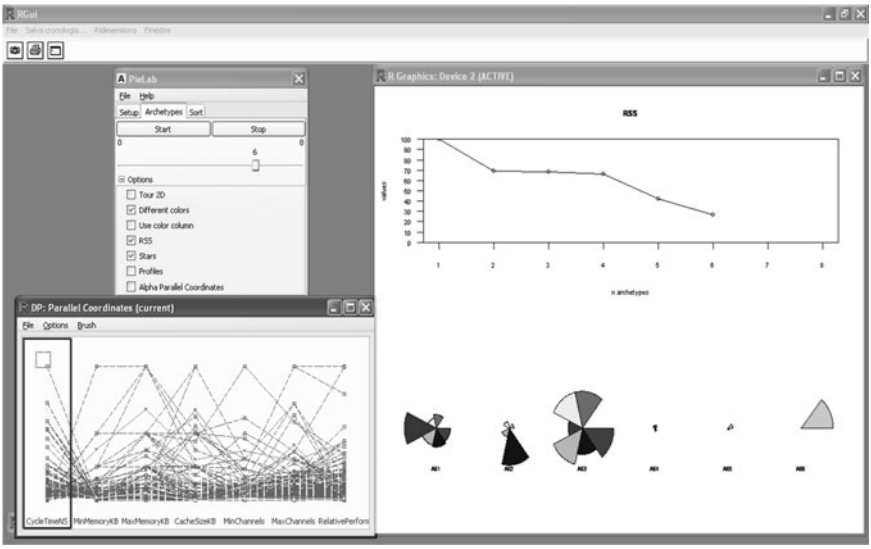


Fig. 3 A spreadplot for the CPUs dataset for $m = 6$ archetypes. *Clockwise*: the control panel which clearly shows the sliding cursor for interactively choosing the number of archetypes, and the list of possible graphical representations; the $RSS(m)$ function and the star plots for the chosen six archetypes; the parallel coordinate plot of the CPUs data along with the archetypes

4.2 Representing Data in the Spaces Spanned by the Archetypes

The next step of the proposed procedure relies on the representation of the data in the m-dimensional spaces spanned by the archetypes.

The archetypes are vertices of a simplex in the data space $\Re^p$, and for each data point $\mathbf{x}'_i$ new coordinates with respect to the archetypes can be obtained by solving the equation $(\lambda_{i1} + \cdots + \lambda_{im})\,\mathbf{x}'_i = \lambda_{i1}\,\mathbf{a}'_1 + \cdots + \lambda_{im}\,\mathbf{a}'_m$.

The coefficients $(\lambda_{i1}, \ldots, \lambda_{im})$ are the new coordinates of $\mathbf{x}'_i$ in an associated space, and they are called barycentric coordinates (see e.g. Coxeter 1969) with respect to the archetypes. The archetypes themselves have barycentric coordinates $(1, 0, \ldots, 0), (0, 1, \ldots, 0), \ldots, (0, 0, \ldots, 1)$, as they are the associated space basis. We note that, from the geometric properties of the barycentric coordinates, in the space spanned by the archetypes the data points actually belong to an $(m - 1)$ dimensional subspace of this associated space.

The reconstructed data points $\tilde{\mathbf{x}}'_i$ have barycentric coordinates in the archetype associated space as well. In particular, the equation $(\lambda_{i1} + \cdots + \lambda_{im})\,\tilde{\mathbf{x}}'_i = \lambda_{i1}\,\mathbf{a}'_1 + \cdots + \lambda_{im}\,\mathbf{a}'_m$, is exactly solved for $\lambda_{ij} = \gamma_{ij}$, $j = 1, \ldots, m$. Hence, it turns out that the $\gamma_{ij}(m)$ coefficients are the barycentric coordinates for the reconstructed points in the associated space (see Porzio et al. 2008 for details).

Consequently, they may be exploited to map the original data into a lower dimensional space. Given the relationship between each $\mathbf{x}'_i$ and its corresponding $\tilde{\mathbf{x}}'_i$, each

original point will be represented by the coefficients γ_{ij} into the archetype associates space. Note that the same dataset may be mapped into many archetype spaces, one for each value of m. Analyzing data in these associated spaces provides further insights into the data structure from a different perspective.

4.3 Exploring the Peripheries of the Data Scatter

As previously stated, when the aim is to explore the data, it can be useful to look at the archetypes for different values of m. Indeed, it is worth noticing that looking at a set of values of m may be necessary – even if the $RSS(m)$ indicates that a small m is sufficient to well reconstruct the data – to search for outliers and peripheral groups. In such cases, it could happen that outliers will coincide with some archetypes not corresponding to a sharp decrease in the $RSS(m)$. Indeed, the first archetypes will catch the majority of data, while additional archetypes will point to outliers or to small peripheral groups, if any.

The exploration task can be pursued exploiting static and dynamic graphical representation, such as parallel coordinates and tourplot, in the associated space provided by the barycentric coordinates. In the parallel coordinate plot of such associated spaces, multivariate skewness could be highlighted as marginal asymmetry along some directions. Moreover interaction tools – brushing, slicing and coloring – can offer the user a thorough exploration of the data periphery. Brushing and coloring data around the archetypes in the parallel coordinate plot can highlight gaps in the data structures, small groups and outliers. To enhance the detection of interesting patterns the parallel coordinate plot can also be empirically linked to the corresponding data in the tourplot graph.

For example, in Fig. 4, which correspond to $m = 4$ archetypes, the parallel coordinate plot depicts all the data in a 4-D space where all the coordinates are the γ coefficients as previously stated (for the geometric properties of γ coefficients it is actually a 3-D space). The plot shows that, for three out of the four axes, some points are isolated even if not very far from the others, while for the $a_3(4)$ axis there are three data groups. By pointing to the observations close to $a_3(4)$ – i.e. those points with coordinate values close to the pattern (0, 0, 1, 0) – we highlight three data points in the tourplot along the Cycle time dimension. Further investigations can be pursued to analyze the other groups appearing on the same archetype-axis. Some skewness and clusters appear by looking along the directions of the first two archetypes.

While in Sect. 4.1 we observed that going from $m = 4$ to $m = 6$ there was not much gain in interpretability and synthesizing power, when exploring the peripheries of the data scatter, the analysis is enhanced at $m = 6$ archetypes. In fact in Fig. 5, more peripheral groups appear: in four out of six dimensions small peripheral groups can be detected. For example, by pointing to the observations close to $a_3(6)$ – i.e. those points with coordinate values close to the pattern (0, 0, 1, 0, 0, 0) – we highlight in the tourplot twelve isolated data points along the Cycle time

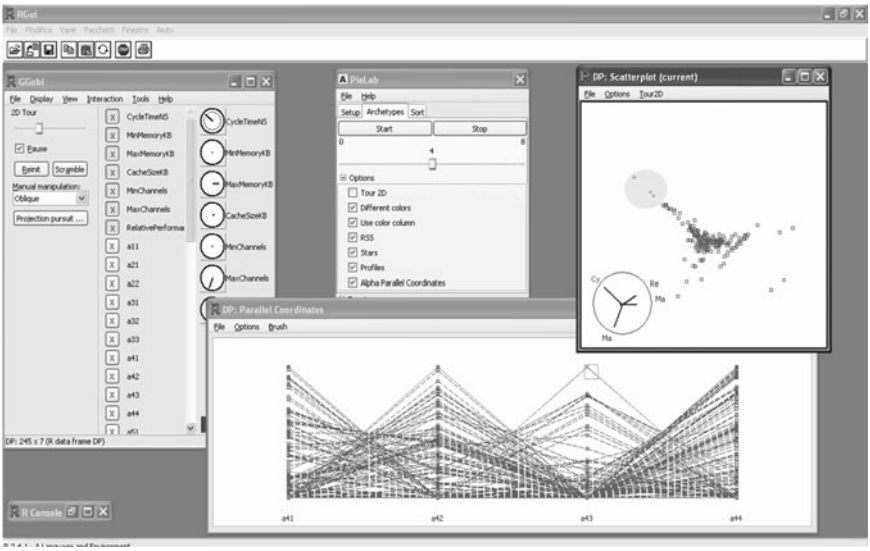


Fig. 4 The strategy in action for $m = 4$ archetypes: acting through brushing and coloring in the parallel coordinate plot in the γ space involves the highlighting of the corresponding points in the tourplot

dimension. Remember that with $m = 4$ archetypes we succeeded in detecting only three isolated data points. Furthermore, the first three archetypes and the last one show severe skewness. It is worth noticing that the tourplot view in Fig. 5 spotting the isolated group is not straightforward. In our case, with the aid of data representation in the space spanned by the archetypes and of interactive graphics, it was easier for us to capture this particular projection.

5 Concluding Remarks

The approach we have used in this paper to explore multidimensional data scatter seems promising in "finding and revealing the clues", mainly in uncovering gaps in the data structures, small groups, outlying values, asymmetries and irregularities in the shape.

The whole procedure, being based on open source softwares, can be easily replicated and perhaps enhanced. Moreover it is not computationally intensive and it does not require huge hardware capabilities.

Finally, note that for the sake of presentation we split the procedure into several spreadplots. However, all the different views can be merged in a unique spreadplot and inspected at the same time.

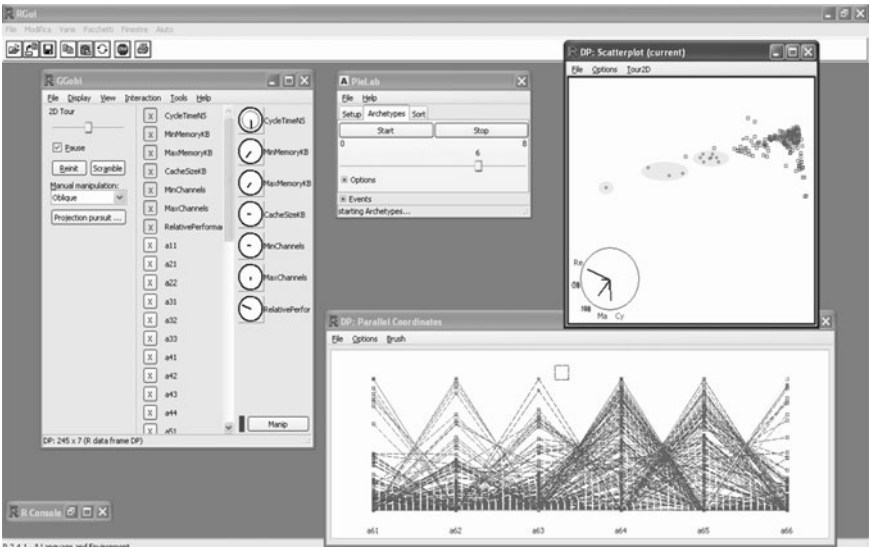


Fig. 5 The strategy in action for $m = 6$ archetypes: acting through brushing on the parallel coordinate plot in the γ space involves the highlighting of the corresponding points in the tourplot

Acknowledgements The authors wish to thank Michele Risi from University of Salerno for the design and implementation of the software architecture. The research work of Maria Rosaria D'Esposito benefits from the research structures of the STATLAB at the Department of Economics and Statistics, University of Salerno. The research work of Domenico Vistocco is supported by Laboratorio di Calcolo ed Analisi Quantitative, Department of Economics, University of Cassino.

References

Buja, A., Lang, D. T., & Swayne, D. F. (2003). GGobi: Evolving from XGobi into an extensible framework for interactive data visualization. *Journal of Computational Statistics and Data Analysis, 43*, 423–444.

Buschmann, F., Meunier, R., Rohnert, H., Sommerlad, P., & Stal, M. (1996). *A system of patterns: Pattern-oriented software architecture*. West Sussex, England: Wiley.

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical methods for data analysis*. Wadsworth, Monterey CA.

Chan, B. H. P., Mitchell, D. A., & Cram, L. E. (2003). Archetypal analysis of galaxy spectra. *Monthly Notice of the Royal Astronomical Society, 338*, 790–795.

Cook, D., & Swayne, D. F. (2007). *Interactive and dynamic graphics for data analysis: With R and GGobi*. Berlin, Heidelberg: Springer, Use R Series.

Coxeter, H. S. M. (1969). *Introduction to geometry* (2nd ed., pp. 216–221). §13.7, Barycentric coordinates. New York: Wiley.

Cutler, A., & Breiman, L. (1994). Archetypal analysis. *Technometrics, 36*, 338–347.

D'Esposito, M. R., & Ragozini, G. (2008). A new R-ordering procedure to rank multivariate performances. In *Quaderni di Statistica* (Vol. 10, pp. 5–21). Italy: Liguori Editore.

D'Esposito, M. R., Palumbo, F., & Ragozini, G. (2006). Archetypal analysis for interval data in marketing research. *Italian Journal of Applied Statistics, 18*, 343–358.

Ein-Dor, P., & Feldmesser, J. (1987). Attributes of the performance of central processing units: A relative performance prediction model. *Communications of the ACM, 30*, 308–317.

Elder, A., & Pinnel, J. (2003). Archetypal analysis: An alternative approach to finding defining segments. In *2003 Sawtooth Software Conference Proceedings* (pp. 113–129). Sequim, WA.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1995). *Design patterns: Elements of reusable object-oriented software*. Reading Mass.: Addison Wesley.

Heavlin, W. D. (2007). Archetypal analysis of computer performance. *38th Symposium on the INTERFACE on Massive Data Sets and Stream*, Pasedena, CA.

Lang, D. T., Swayne, D., Wickham, H., & Lawrence, M. (2008). *rggobi: Interface between R and GGobi*. R package version 2.1.10, URL http://www.ggobi.org/rggobi.

Li, S., Wang, P., Louviere, J., & Carson, R. (2003). Archetypal analysis: A new way to segment markets based on extreme individuals. *A Celebration of Ehrenberg and Bass: Marketing Knowledge, Discoveries and Contribution, ANZMAC 2003 Conference Proceedings* (pp. 1674–1679). Adelaide.

Marinetti, S., Finesso, L., & Marsilio, E. (2006). Matrix factorization methods: Application to thermal NDT/E. *NDT&E International, 39*, 611–616.

Marinetti, S., Finesso, L., & Marsilio, E. (2007). Archetypes and principal components of an IR image sequence. *Infrared Physics & Technology, 49*, 272–276.

Porzio, G. C., Ragozini, G., & Vistocco, D. (2008). On the use of archetypes as benchmarks. *Applied Stochastic Models in Business and Industry, 24*, 419–437.

R Development Core Team. (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Stone, E. (2002). Exploring archetypal dynamics of pattern formation in cellular flames. *Physica D, 161*, 163–186.

Stone, E., & Cutler, A. (1996). Introduction to archetypal analysis of spatio-temporal dynamics. *Physica D, 96*, 110–131.

Stuetzle, W. (1987). Plot windows. *Journal of American Statistical Association, 82*, 466–475.

Tukey, J. (1977). *Exploratory Data Analysis*. Reading, MA, U.S.A.: Addison-Wesley.

Wegman, E. J., & Carr, D. B. (1993). Statistical graphics and visualization. In C. R. Rao (Ed.), *Handbook of Statistics: Computational Statistics* (Vol. 9, pp. 857–958). NY, U.S.A.: Elsevier, North Holland.

Wilhelm, A. (2005). Interactive statistical graphics: The paradigm of linked views. In C. R. Rao, E. J. Wegman and J. L. Solka (Eds.), *Data mining and data visualization* (Vol. 24 of Handbook of Statistics, pp. 437–537). NY, U.S.A.: Elsevier, North-Holland.

Yang, D., Rundensteiner, E. A., & Ward, M. O. (2007). Analysis guided visual exploration to multivariate data. *IEEE Symposium on Visual Analytics Science and Technology*. Sacramento, California.

Young, F. W., Faldowski, R. A., & Harris, D. F. (1992). The spreadplot: A graphical spreadsheet of algebraic linked of dynamic plot. In: *ASA Proceedings Section on Statistical Graphics*. American Statistical Association, Alexandria, VA.

Young, F. W., Faldowski, R. A., McFarlane, M. M. (1993). Multivariate statistical visualization. In C. R. Rao (Ed.), *Handbook of Statistics: Computational Statistics* (Vol. 9, pp. 959–998). NY, U.S.A.: Elsevier, North Holland.

Exploring Sensitive Topics: Sensitivity, Jeopardy, and Cheating

Claudia Becker

Abstract The general problem of nonresponse or false reporting in surveys is well known. Especially when asking questions about so-called sensitive topics, these effects can be severe. One proposal to overcome the problem is to use randomized response techniques for reducing bias due to these effects. The original work of Warner (1965) deals with estimating the proportion of people in a population having a certain (sensitive) characteristic. Later, also the collection of quantitative sensitive variables came into focus. One general assumption of the models is that people asked with these techniques respond honestly. More recently, researchers in the field realized that also under randomized response we find false reporting due to lying or not understanding the procedures. The paper reviews possible approaches and discusses challenges in exploring sensitive topics and in understanding people's reactions on them.

1 Introduction

Avoiding nonresponse and false reporting is a major challenge when performing questionnaire based surveys. Especially for sensitive topics, people might not want to answer or, if responding, might not answer honestly due to the fear of compromising themselves. To overcome this problem, Warner (1965) proposes to integrate a randomization technique when asking sensitive questions to protect each single respondent's privacy while still being able to conclude on a population. Warner's original work deals with estimating proportions (assigned to binary variables, transferred to the case of qualitative sensitive characteristics by Horvitz et al. 1967; Greenberg et al. 1969), but his approach can also be generalized to situations with quantitative sensitive characteristics (e.g. Greenberg et al. 1971; Eichhorn and Hayre 1983).

C. Becker
School of Law, Economics, and Business, Martin-Luther-University Halle-Wittenberg, D-06099 Halle, Germany
e-mail: claudia.becker@wiwi.uni-halle.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_32,

The general assumption of the models, namely that people respond honestly, even to sensitive questions, because their privacy is protected by the techniques, might be doubted. Even with randomized response, cheating happens. Reasons for this can be seen in the context of the respondents' jeopardy (Leysieffer and Warner 1976), or of the questions' sensitivity level (Gupta et al. 2002; Huang 2004). First approaches in the literature develop procedures to estimate the level of false reporting (Clark and Desharnais 1998). Also, estimation of the topics' sensitivity level is used to judge the reliability of answers.

Whether a topic is sensitive usually depends on the socio-cultural context. For example, in the mid-seventies (Krotki and Fox 1974) investigate women's fertility. In their study, among the sensitive topics we find questions about abortions, sexual activity before marriage, children born illegitimate, or the use of contraceptives. Most of these topics would probably not found to be sensitive nowadays, at least in Europe, since people's attitudes towards them have changed. Hence, the classification of topics according to their current sensitivity – apart from common sense or scientific knowledge – might be helpful.

The paper is organized as follows. In Sect. 2, a brief introduction to randomized response is given. Section 3 is dedicated to approaches for exploring the sensitivity of topics, while Sect. 4 illustrates some problems in the estimation of sensitivity itself. Finally, in Sect. 5 some ideas for further research are given.

2 Randomized Response

For understanding the concept of randomized response, consider the following example for the case of a binary sensitive characteristic. The question of interest is the consumption of illegal drugs during the last six months. Instead of asking respondents directly, "Did you ever, during the last six months, consume illegal drugs?" (with possible answers YES and NO), the interviewees are requested to act according to the following instructions. The respondents get two dice, one red, one white, and are asked to throw the dice and add the resulting numbers. Depending on the result, they have to act further. If they got a total of 4–10 points, then they have to answer YES or NO to the question given above. If, on the other hand, they got a total of less than 4 or more than 10 points, they have to answer YES or NO to the question "Did the red dice show a 2 or 5?" To reach the promised privacy protection, the respondents must not be observed by the interviewer while performing the dice experiment, and the respondents must not give any other answer than YES or NO. Especially, they are not to tell the interviewer to which of the questions the given answer belongs.

The general idea of randomized response techniques for qualitative characteristics is hence to let the respondent answer to the sensitive question with some given probability p, $0 < p < 1$ and with $1 - p$ let him or her answer to something different, where only the respondent knows to which question the answer is given, while the interviewer knows p. "Something different" varies with the various versions of

randomized response. In the example given above, the so-called unrelated question model (Horvitz et al. 1967; Abul-Ela et al. 1967; Greenberg et al. 1969), the alternative question is a question about some topic which has nothing to do with the sensitive characteristic.

For quantitative characteristics, a similar approach is also possible (see Greenberg et al. 1971), but here it is more convenient to let the respondent report a scrambled answer (e.g. by multiplying the value of his/her sensitive characteristic by some random number, cf. Eichhorn and Hayre 1983), where again only the respondent knows the true value and the interviewer knows the distribution of the scrambling process.

Compared to asking sensitive questions directly, the randomized response approaches have the advantage that the answers are given anonymously, since the interviewer does not know which question is answered (or which is the true quantitative value). Hence, there is no conclusion possible about the individual respondent, but due to applying the method to a whole sample, still information about a population can be generated by the according estimators for the proportion of people carrying the sensitive characteristic (in the qualitative case), or for the population mean (in the quantitative case), respectively. The hopes which are connected with randomized response are that respondents accept the protection of their privacy by the method and hence do not feel the necessity to report falsely to the sensitive question. When constructing estimators and comparing the various randomized response versions, the general assumption hence is that only honest answers are collected due to the randomization.

3 Exploring Sensitivity

When exploring people's reactions to randomized response, one realizes that the general assumption of honest answers might be questioned. In summer 2007, we conducted an experiment on randomized response using an unrelated question model. The setting (two dice, numbers) was as described at the beginning of Sect. 2, the question asked was "Have you ever, during the last six months, deliberately dodged the fare in public transport?" German associations for public transport estimate from their own experiences of controls in buses, trains etc. that about 13–18% of people using public transport do not pay the fares. The experiment yielded an estimated proportion of about 28% traveling without paying the fare (females: 20%, males: 35%). This surprisingly high value at seems to be a mixture of two effects. On the one hand, the use of randomized response contributes to the increased value (compared to the impression given by the public transport associations). Experiences from studies show that this desired effect of reducing the estimated number of unreported cases indeed results from randomized response techniques (one example can be seen in the study of Goodstadt and Gruson 1975). On the other hand, in our case the sample was biased towards younger people, since the experiment was conducted during university's open house, where traditionally many pupils and students

show up. It might be assumed that the amount of persons dodging the fare is higher among younger people.

One necessity for getting honest answers in such randomized response settings is that people trust the protection of their privacy given by the randomization. Already in the work of Leysieffer and Warner (1976) it is mentioned that this protection might not be given if the randomization mechanism is not constructed with the necessary precision. They define the so-called jeopardy, meaning that a respondent is jeopardized if for a binary sensitive variable the probability

$$\begin{aligned} &P(\text{belonging to sensitive group A}|\text{answer induced by randomization}) \\ &> P(\text{belonging to A}) \end{aligned}$$

which results in a decreasing willingness to cooperate (an hence an increase in nonresponse), since the respondent feels exposed rather than protected. Leysieffer and Warner conclude that the question design has to be chosen accordingly to decrease jeopardy. Whether jeopardy is really a problem obviously depends on whether the respondent realizes the probabilities mentioned above.

To supplement the results of our experiment, we also asked the respondents whether they thought that their privacy was really protected by the method ("Do you think that now we really do not know whether you personally dodged the fare?"). Here, about 59% (females: 57%, males: 61%) signalled their trust, meaning that more than 40% of the respondents did not assume that their privacy was really guaranteed by the method. Hence, we conclude that a lack of trust into the method exists which may cause problems when applying such randomized response techniques.

Clark and Desharnais (1998) explicitly assume that there is a certain amount of respondents not reacting according to the randomized response instructions, the so-called cheaters. Clark and Desharnais introduce a special version of randomized response for a dichotomous sensitive characteristic, the forced answer randomized response, to estimate the proportion of cheaters in a sample. In this randomized response version, the respondents, with probability p, have to answer to the interesting question ("Do you belong to the sensitive group?"), with $1 - p$ they are just told to answer "Yes" (independently of their membership to the sensitive group). Since in this setting, the answer "Yes" is potentially stigmatizing, Clark and Desharnais assume that cheaters always answer "No". Then in their model it is possible to estimate the proportion of cheaters and to test this proportion against zero. Clark and Desharnais do not propose any adjustment of the randomized response estimator, but conclude that, if the amount of cheaters is too large, the survey results are not to be used. In their work it is not discussed whether the cheater assumption (cheaters always answer "No") is realistic. Moreover, the problem of people not understanding the randomized response instructions at all is not considered.

In our experiment the interviewers had to write down their impression of whether the respondents understood what they had to do according to the instructions presented to them. Although this is only a subjective personal impression, it is quite interesting that according to the interviewers' opinion, only about 82% (females: 80%, males: 84%) of the respondents did really understand the instructions.

Altogether the results of the experiment hint to the fact that a not too small amount of people do not trust the randomized response idea, or do not understand the instructions to perform a randomized response technique as intended. Both can yield the problem of people not answering accordingly to the randomized response type questions, either intentionally, or unintentionally. But the assumption of answering as instructed is still a basic one in randomized response techniques. In the following section, we will briefly introduce the sensitivity level and its estimation and show how this estimation is influenced by people not answering according to the instructions.

4 The Sensitivity Level

Gupta et al. (2002) introduce the sensitivity level of a topic in the framework of a quantitative sensitive variable. They define the method of optional randomized response: the respondent chooses to either answer directly, or to give a scrambled answer according to some randomization mechanism. The sensitive question asked is of the type "Which value does the interesting variable have in your case?" (e.g. "What amount of your income did you not declare for the tax declaration last year?"). The sensitivity level of the topic is then defined as the probability that scrambling is chosen instead of direct answering.

Denote, more formally, the sensitive variable by $X > 0$, and let the scrambling variable be $S > 0$, then the respondents report $Y = S^Z X$, where $Z = 1$ if the answer is scrambled, $Z = 0$ otherwise. The sensitivity level is given as $E(Z) = w$. Gupta et al. derive estimators for the mean μ_X of the sensitive variable X as well as for the sensitivity level w.

In this model the central assumptions are that w is a population parameter (especially it is not varying individually between the respondents), w is independent of X, and all respondents understand the randomized response instructions and answer honestly according to these instructions.

In the results of our experiment (dodging the fare in public transport) we see that there exists a nonignorable amount of non-understanders and of non-trusting respondents. Moreover, it can also be shown that at least in this experiment trust and understanding are independent. Hence the assumption of people understanding and acting according to the instructions might be doubted. Finally, it can be noted that some people, independently of what they were instructed to do, openly reported that they did never dodge the fare intentionally, while none of the respondents openly reported having dodged the fare. Hence, it can also be concluded that the assumption of w being independent of X might not hold.

A small simulation shows the influence of violating these assumptions (w independent of X, all respondents acting according to instructions) on the estimation of μ_X and w. Assume the setting given in Gupta et al. (2002), where $X = \alpha + \beta V + V^{1/2}\varepsilon$, with $V \sim \Gamma(1/1.44, 5)$ (shape $= 1/1.44$, rate $= 1/5$), $\varepsilon \sim N(0, 1)$, $S \sim \chi^2_1$. The values of α and β are chosen yielding $\mu_X = 24.167$. Scrambling

Table 1 Simulation results for sensitivity level: estimated values when violating the assumptions

	$c = 90\%$-quantile			$c = 80\%$-quantile		
n	w	$\widehat{w}$	$\widehat{\mu}_X$	w	$\widehat{w}$	$\widehat{\mu}_X$
100	0.5	0.481	23.167	0.5	0.462	22.618
100	0.6	0.574	23.165	0.6	0.555	22.625
100	0.9	0.859	23.144	0.9	0.825	22.600
500	0.5	0.483	23.155	0.5	0.465	22.595
500	0.6	0.579	23.159	0.6	0.557	22.612
500	0.9	0.863	23.171	0.9	0.833	22.601
1,000	0.5	0.484	23.167	0.5	0.466	22.606
1,000	0.6	0.579	23.167	0.6	0.557	22.612
1,000	0.9	0.864	23.158	0.9	0.833	22.592

is simulated as described below, for each respondent $Y = S^Z X$ is recorded. The simulation is run with 20,000 replications for samples of sizes $n = 100, 500, 1,000$ and sensitivity levels $w = 0.5, 0.6, 0.9$.

To model the dependency of w on X, the respondents' behavior is modelled in the following way: if $X <$ cut-off value c, then the sensitivity level w is identical for all respondents, and hence a proportion w of the respondents reports the scrambled value of X, while $1-w$ report their true value. If, on the other hand, $X \geq c$, then all respondents report only half of their true value. The value c is chosen as a certain quantile of the X distribution. The results are summarized in the following Table 1.

Obviously, violating the assumptions influences the estimation quality. In general, we can see a tendency of the sensitivity level being slightly underestimated (getting worse for larger sensitivity levels), while the influence is stronger for estimating the population mean of the sensitive characteristic (which is also underestimated). The effects become better visible for larger amounts of people not acting according to the instructions (c chosen as the 80% quantile), and it may be conjectured that we can generate even stronger effects if choosing a different answering mechanism for the non-compliers. Moreover, the effects do not vanish for larger sample sizes.

5 Conclusions

Further research on the application of randomized response techniques will have to deal with the fact that the strong assumptions like people acting according to instructions given to them and answering honestly due to the protecting effect of the methods do not hold in practice. In the special case of estimating the sensitivity level of a certain topic it can be seen that already small deviations from these assumptions disturb the results clearly. Moreover, some of the assumptions might also have to be questioned, e.g. the assumption of the sensitivity level being independent of the values of the sensitive variable and identical for the whole population.

Hence, it might be necessary to find ways to define the sensitivity level differently. For modelling answering behavior, it might be interesting to integrate behavioral science into models and estimators. Also, robust estimation of μ_X and w could be a promising approach to overcome the problems connected with peoples' behavior in reality.

References

Abul-Ela, A. L. A., Greenberg, B. G., & Horvitz, D. G. (1967). A multi-proportions randomized response model. *Journal of the American Statistical Association, 62*, 990–1008.

Clark, S. J., & Desharnais, R. A. (1998). Honest answers to embarrassing questions: Detecting cheating in the randomized response model. *Psychological Methods, 3*, 160–168.

Eichhorn, B. H., & Hayre, L. S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference, 7*, 306–316.

Goodstadt, M. S., & Gruson, V. (1975). The randomized response technique: A test on drug use. *Journal of the American Statistical Association, 70*, 814–818.

Greenberg, B. G., Abul-Ela, A. L. A., Simmons, W. R., & Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association, 64*, 520–539.

Greenberg, B. G., Kuebler, R. R., Abernathy, J. R., & Horvitz, D. G. (1971). Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association, 66*, 243–250.

Gupta, S., Gupta, B., & Singh, S. (2002). Estimation of sensitivity level of personal interview survey questions. *Journal of Statistical Planning and Inference, 100*, 239–247.

Horvitz, D. G., Shah, B. V., & Simmons, W. H. (1967). The unrelated question randomized response model. *Proceedings of the Social Statistics Section* (pp. 65–72). American Statistical Association.

Huang, K. C. (2004). A survey technique for estimating the proportion and sensitivity in a dichotomous finite population. *Statistica Neerlandica, 58*, 75–82.

Krotki, K., & Fox, B. (1974). The randomized response technique, the interview, and the self-administered questionnaire: An empirical comparison of fertility reports. *Proceedings of the Social Statistics Section* (pp. 367–371). American Statistical Association.

Leysieffer, F. W., & Warner, S. L. (1976). Respondent Jeopardy and optimal designs in randomized response models. *Journal of the American Statistical Association, 71*, 649–656.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60*, 63–69.

Sampling the Join of Streams

Raphaël Féraud, Fabrice Clérot, and Pascal Gouzien

Abstract One of the most critical operators for a Data Stream Management System is the join operator. Unfortunately, the join operator between the stream A and B is a blocking operator: for each current tuple of the stream A, the entire stream B have to be scanned. The usual technique used for unblocking stream operators consists to restrict the processing to a sliding window. This technique emphasizes recent data which are considered to be more relevant than old data. However, in a Data Stream Management System, a general approach is needed to join any data streams for any applications. Our approach is to consider data stream join as an estimation problem. The estimation model is simple and generic: a reservoir per data stream is used to model the join. The quality of join estimator is based on the frequencies of join key in the join. We propose four algorithms to feed reservoirs. The proposed methods outperform reservoir sampling approach on synthetic and real data streams.

1 Introduction

For a telecommunication operator there is at least two major applications of stream mining techniques:

- The first one consists of processing network sensor. To build indicators for monitoring the network, sensors are installed on network nodes. Each sensor is equiped with processors and memory. Each sensor emits a stream. The monitoring module joins and aggregates data from each distributed stream.
- The second major application for a telecommunication operator is the feeding of information system. Each application (services, CRM, Billing, etc.) produces data. These data are written on operational data stores. Each day, the operational data stores are used to feed the datawarehouse. To feed datawarehouse

R. Féraud (✉)
Raphaël Féraud, Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion, France
e-mail: raphael.feraud@orange-ftgroup.com

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_33,

at lower cost, tables can be processed as streams on a centralized Extraction Transformation and Loading server.

These two applications lead to different constraints. For processing network sensor, each stream has to produce its result without information about others streams. For feeding information system, streams are processed together on a centralized server. Therefore streams can exchange information. In the next section we propose a general framework for sampling join of streams. Four algorithms using this framework are detailed in the following section. These algorthms are tested first on a synthetic problem and then on a real problem. On the last part we achieve the comparison with a reminder of application constraints.

2 General Framework

Our goal is to be able to answer various queries on the join of two data streams F_1 and F_2 at any point t in time: Unfortunately, the join operation is blocking. The join cannot be emitted as a flow without keeping F_1 and F_2 in memory under finite memory constraint, the whole join cannot be produced (see Babcock et al. 2002 for an overview of stream mining). The join must be estimated from samples of F_1 and F_2. To produce and maintains a sample from a stream of size n, Reservoir Sampling algorithm (Vitter 1985) an inclusion probability of $n/(t+1)$ is given for each tuple arrived at time t. An interesting property of this algorithm is that, when t tuples have been observed, all the t tuple have the same probability to be included in the reservoir: n/t. It exists biased version of this algorithm, to take into account recent data (Aggarwal 2006) or weighted data (Chaudhuri and Motwani 1999; Kolonko and Wasch 2004; Efraimidis and Spirakis 2004). A first solution for Sampling join consists of sampling each stream using reservoir sampling and then join them. However, the drawback of this approach is that when the number of tuples is important each reservoir can contain a lot of tuples that do not join. In the stream mining framework, memory resources are limited, and it is crucial to avoid the wasting space. Chaudhuri and Motwani (1999) and Chaudhuri et al. (1999), proposed an algorithm to avoid the wasting space. A first step consists to draw a biased sample for the first stream with weight corresponding to the frequency on the second stream. The second step consists to complete the sample with tuples of the second stream. Another interesting approach is the Ripple join algorithm (see Hellerstein et al. 1997; Hellerstein and Haas 1999; Hellerstein et al. 2000). The idea is to produce at anytime an estimation of the query with a confidence interval. The user stops the query when the confidence interval is sufficient. The drawback of this approach is that the computational time and the memory resources needed are unknown in advance. A general framework is proposed by Das et al. (2003). The model is composed by a reservoir for each stream. When a query is requested, reservoirs are joined. The model can be integrated to exchange information between each reservoir, or modular to be decentralized. We use and develop this general framework for sampling joins of streams.

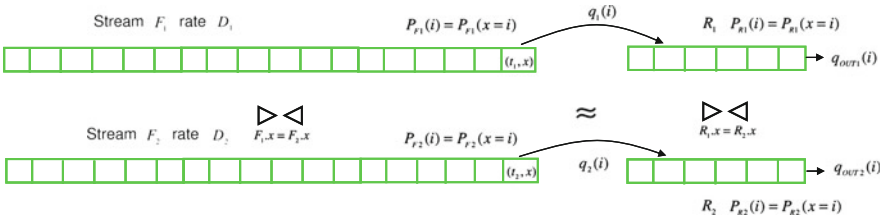


Fig. 1 General framework

The sample of the join consists in two or more reservoirs which can be joined at any time. We have to consider four probabilities by stream (see Fig. 1):

- $P_F(i)$, the probability of join key i in the stream F,
- $P_R(i)$, the probability of join key i in the reservoir R,
- $q(i)$, the inclusion probability of join key i in the reservoir,
- $q_out(i)$, the exclusion probability of join key i from the reservoir.

3 Four Algorithms for Sampling the Join of Streams

3.1 Reservoir Sampling

To draw a sample of size n from a stream, reservoir sampling algorithm (Vitter 1985) takes the n first points. Then, each following tuples randomly replace one of the tuples of the reservoir with the probability:

$$q(t) = |R|/(t+1)$$

It can be shown that when t tuples have been processed, the probability of each tuple of the reservoir is $|R|/t$. Reservoir sampling allows to draw efficiently an uniform sample for each stream, which can be joined at any time.

3.2 Weighted Reservoir Sampling

The drawback of reservoir sampling is that each reservoir may contain a lot of keys which do not join. When memory resources are limited, it is crucial to avoid this wasting reservoir space. In other words, we need to reach in each reservoir a probability distribution of each key which maximizes the size of the join and which respects the key distribution in the join with a bounded reservoir size. If we formalize our idea we obtain that we want to maximize the join size:

$$|R_1| \cdot |R_2| \cdot \sum_j P_{R_1}(j) \cdot P_{R_2}(j)$$

under three constraints:

- $\sum_j P_{R_1}(j) = 1$ and $\forall i \ P_{R_1}(i) > 0$, $\sum_j P_{R_2}(j) = 1$ and $\forall i \ P_{R_2}(i) > 0$, P_{R_1} and P_{R_2} are probabilities,
- $R_1 + R_2 = R$, the sum of the size of each reservoir is bounded,
- $\frac{P_{R_1}(i).P_{R_2}(i)}{\sum_j P_{R_1}(j).P_{R_2}(j)} = P_{F_1F_2}(i)$, there is no sampling bias,

The previous optimisation problem is equivalent to maximize independently:

- $|R_1| \cdot |R_2|$ under the constraint $R_1 + R_2 = R$
- $\sum_j P_{R_1}(j).P_{R_2}(j)$ under the constraints $\sum_j P_{R_1}(j) = 1$ and $\forall i \ P_{R_1}(i) > 0$, $\sum_j P_{R_2}(j) = 1$ and $\forall i \ P_{R_2}(i) > 0$

Therefore, the optimal solution is:

- $R_1 = R_2 = R/2$,
- $\forall i \ P_{R_1}(i) = P_{R_2}(i)$ and as $P_{F_1F_2}(i) = \frac{P_{F_1}(i).P_{F_2}(i)}{\sum_j P_{F_1}(j).P_{F_2}(j)}$ then the optimal solution is obtained using: $P_{R_1}(i) = P_{R_1}(i) \propto \sqrt{P_{F_1F_2}(i)}$

It can be shown that the obtained join size $|J|$ is larger than the one obtained using Reservoir Sampling. Using Reservoir Sampling, with two reservoirs of equal size $R/2$, we have:

$$\overline{|J_{RS}|} = \left(\frac{R}{2}\right)^2 \sum_i P_{F_1}(i).P_{F_2}(i)$$

Using Weighted Reservoir Sampling, we have:

$$\overline{|J_{WRS}|} = \left(\frac{R}{2}\right)^2 \sum_i P_{R_1}(i).P_{R_2}(i)$$

$$\Leftrightarrow \overline{|J_{WRS}|} = \left(\frac{R}{2}\right)^2 \frac{\sum_i P_{F_1}(i).P_{F_2}(i)}{\left(\sum_j \sqrt{P_{F_1}(j).P_{F_2}(j)}\right)^2}$$

We have $\left(\sum_j \sqrt{P_{F_1}(j).P_{F_2}(j)}\right)^2 \leq 1$ and $\frac{\sum_i P_{F_1}(i).P_{F_2}(i)}{\left(\sum_j \sqrt{P_{F_1}(j).P_{F_2}(j)}\right)^2} \leq 1$ then

$$\overline{|J_{RS}|} \leq \overline{|J_{WRS}|} \leq \left(\frac{R}{2}\right)^2$$

3.3 Deterministic Reservoir Sampling

To obtain the optimal estimator, the minimum variance estimator, a sample drawn from the join key distribution is needed. It is the purpose of Deterministic Reservoir Sampling. Deterministic Reservoir Sampling uses three steps to reach this goal:

- At the first step a sampling design based on the join key is built.
- At the second step the samples R_1 and R_2 are collected.
- At the third step the obtained samples R_1 and R_2 are optimised.

The sampling design consists of the draw of $|J|$ keys according to the distribution of the join key in the join:

- $J = 0$
- For $i = 1$ to $|J|$ do
  - Draw a key k from $P_{F_1 F_2}(k)$
  - $J = J + k$

The second step consists of the collect of the J desired keys. Streams are read until size of desired join is obtained. Keys are included for both reservoirs in order to respect the sample design for each key:

- $R_1 = 0$
- $R_2 = 0$
- While $|R_1| \cdot |R_2| < |J|$ do
  - If a key k arrive from F_1 then if $|R_1(k)|<|J(k)|$ and $|R_1(k)|\cdot \max(1, |R_2(k)|) < |J(k)|$ then $R_1 = R_1 + k$
  - If a key k arrive from F_2 then if $|R_2(k)|<|J(k)|$ and $|R_2(k)|\cdot \max(1, |R_1(k)|) < |J(k)|$ then $R_2 = R_2 + k$

The purpose of the last step is to avoid rounding error between the obtained samples R1 and R2 and the sampling design J. For each key, stop condition of the second step is that the number of obtained values is greater than the desired values. For large values of $|J|$, the rounding error can be important. For example, with a sample design of $1,001$ items for a particular key, if 1,000 items are obtained on R_1 and 2 on R_2, the rounding error is on the same order of the number of items needed. The last step of Deterministic Sampling Reservoir allows to reduce this rounding error. It is done only when a query is requested on the join.

- For all join key i in $|R_1 >< R_2|$
  - $E_1 = (|R_2(i)| \cdot (|R_1(i)| - 1) - |J(i)|))^2$
  - $E_2 = (|R_1(i)| \cdot (|R_2(i)| - 1) - |J(i)|))^2$
  - $E = (|R_2(i)| \cdot |R_1(i)| - |J(i)|))^2$
- While ($E_1 < E$ and $|R_1(i)| > 1$) or ($E_2 < E$ and $E_2 < E_1$ and $|R_2(i)| > 1$)
  - if ($E_1 < E_2$) then exclude a key i from R_1
  - if ($E_2 < E_1$) then exclude a key i from R_2
  - Evaluate E, E_1, E_2

3.4 Active Reservoir Sampling

We want a sample where the distribution of the join key is as close as possible of the true one. Here, the idea is to minimize the Khi^2 between $P_{R_1R_2}$ and $P_{F_1F_2}$:

$$Khi^2(P_{R_1R_2}, P_{F_1F_2}) = |R_1 >< R_2| \sum_i \frac{\left(P_{R_1R_2}(i) - P_{F_1F_2}(i)\right)^2}{P_{F_1F_2}(i)}$$

The optimisation is done by controlling inclusion probabilities $q_1(i)$ and $q_2(i)$:

$$q_1^{t+1} = q_1^t - \alpha \frac{\partial Khi^2(P_{R_1R_2}, P_{F_1F_2})}{\partial q_1(k)} \tag{1}$$

Update equation is obtained by the derivation of the (1):

$$\frac{\partial Khi^2(P_{R_1R_2}, P_{F_1F_2})}{\partial q_1(k)} = \sum_i \frac{|R_1 >< R_2|}{P_{F_1F_2}(i)} \cdot \frac{\partial K(i)}{\partial q_1(k)}$$

The first term of the previous equation is constant for any given key i, and $K(i) = \left(P_{R_1R_2}(i) - P_{F_1F_2}(i)\right)^2$. For the reservoir R_1, we have:

$$\frac{\partial K(i)}{\partial q_1(k)} = \sum_j \frac{\partial K(i)}{\partial P_{R_1}(j)} \cdot \frac{\partial P_{R_1}(j)}{\partial q_1(k)} \tag{2}$$

The first term of the (2) is obtained by direct derivation, and the second term is obtained from the balance equation between inputs and outputs:

$$\frac{\partial P_{R_1}(j)}{\partial q_1(k)} = -(1-\lambda_{jk})\frac{D_1}{|R_1|}P_{F_1}(k)\frac{|R_1(j)|}{|R_1|} + \lambda_{jk}\frac{D_1}{|R_1|}\left(P_{F_1}(k) - P_{F_1}(k)\frac{|R_1(j)|}{|R_1|}\right) \tag{3}$$

Where $\lambda_{ij} = 1$ if $i = j$ and 0 else, and $|D_1|$ is the rate of the stream 1.

4 Experimental Results

The accuracy of the system is given by:

- The variance (robustness) of the estimator,
- The size of the obtained sample join which is linked to the confidence interval on the query result,
- and the memory resources used.

To evaluate each of these indicators, a toy problem is used: two streams, containing three keys with very different probabilities (see Table 1), are joined. Variances

Table 1 A toy problem: frequencies table of each stream

Key	F_1	F_2	$F_1 >< F_2$
1	0.05	0.02	0.012
2	0.01	0.9	0.10
3	0.94	0.08	0.88

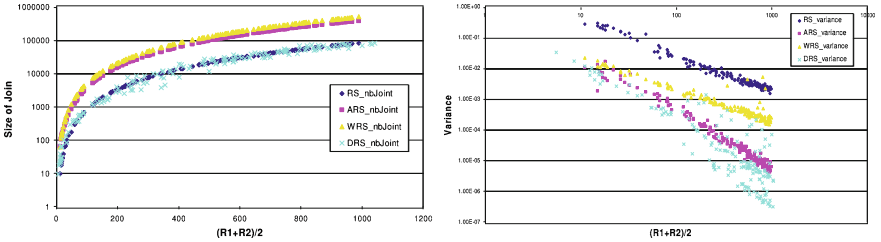


Fig. 2 All the following results are obtained using 100 draws for each value

Table 2 A real problem: frequencies table of each stream

	F_1	F_2	$F_1 >< F_2$
Size	408333	30239496	19550134962
Number of joined keys	74117	61381	61357

and size of joins versus size of reservoirs are observed using 100 draws for each value (see Fig. 2). We observe that Reservoir Sampling estimator is not robust: large variance and small size of join. Deterministic Reservoir Sampling estimator is robust: smaller variance than other estimators. Weighted Reservoir Sampling estimator leads to a large size of join, but with a variance higher than Deterministic Reservoir Sampling. Active Reservoir Sampling seems to outperform other algorithms: low variance, and large size of join for limited resources. To confirm these first results, real traces extracted form information system of Orange is used:

- A services subscription trace F1,
- An use of services trace F2.

The first conclusion of the test on real traces is that with a large number of keys (see Table 2), Active Reservoir Sampling cannot be used (computational time cost is n^3). Reservoir Sampling estimators still not robust: large variance and small size of join. Weighted Reservoir Sampling estimator performs better than Reservoir Sampling. Deterministic Reservoir Sampling estimator outperforms other estimators: low variance, and large size of join for limited resources (see Fig. 3).

5 Conclusion and Future Works

In this preliminary work, we proposed four algorithms to build one of the most critical operators for a Data Stream Management System: the join operator. On real traces, Deterministic Reservoir Sampling outperforms other estimators. It seems to

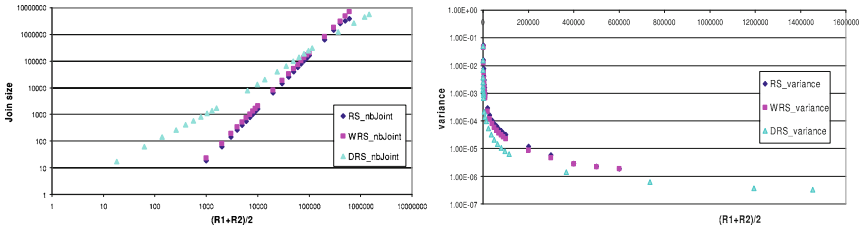


Fig. 3 All the following results are obtained using 100 draws for each value

be the best algorithm for joining streams. Nevertheless, it depends on the application. Deterministic Reservoir Sampling needs to exchange frequencies between each reservoir. For distributed sensors in a telecommunication network, it is not possible. For distributed streams Weighted Reservoir Sampling is well-suited. In a future work, we need to investigate deeper theoretical properties of each algorithm. In particular, we need to bound variances of each estimator and to evaluate the error on different types of queries.

References

Aggarwal, C. (2006). On biased reservoir sampling in the presence of stream evolution. In *VLDB Conference*, 607–618.

Babcock, B., Babu, S., Datar, M., Motwani, R., & Widom, J. (2002). Models and issues in data stream systems. In *ACM SIGMOD*, 1–16.

Chaudhuri, S., & Motwani, R. (1999). On sampling and relational operators. In *IEEE on Data Engineering, 22*, 41–46.

Chaudhuri, S., Motwani, R., & Narasayya, V. (1999). On random sampling over joins. In *ACM SIGMOD*, 263–274.

Das, A., Gehrke, J., & Riedewald, M. (2003). Approximate join processing over data streams. In *ACM SIGMOD*, 40–51.

Efraimidis, P. S., & Spirakis, P. G. (2004). Weighted random sampling. Technical Report *Research Academic Computer Technology Institute*.

Hellerstein, J. M., & Haas, P. J. (1999). Ripple joins for online aggregation. In *ACM SIGMOD*, 287–298.

Hellerstein, J. M., Haas, P. J., & Wang, H. J. (1997). Online aggregation. In *ACM SIGMOD*, 171–182.

Hellerstein, J. M., Avnur, R., & Raman, V. (2000). Informix under CONTROL: online query processing. In *Data Mining and Knowledge Discovery Journal, 4*(4), 281–314.

Kolonko, M., & Wasch, D. (2004). Sequential reservoir sampling with a non-uniform distribution. Technical Report *University of Clausthal*.

Vitter, J. S. (1985). Random sampling with a reservoir. In *ACM SIGMOD, 11*, 37–57.

The R Package fechner for Fechnerian Scaling

Thomas Kiefer, Ali Ünlü, and Ehtibar N. Dzhafarov

Abstract Fechnerian scaling provides a theoretical framework for constructing distances among objects representing subjective dissimilarities. A metric, called Fechnerian, on a set of objects (e.g., colors, symbols, X-ray films, or even statistical models) is computed from the probabilities with which two objects within the set are discriminated from each other by a system (e.g., person, technical device, or even computational algorithm) "perceiving" these objects. This paper presents the package **fechner** for performing Fechnerian scaling of object sets in R. We describe the functions of the package and demonstrate their usage on real Morse code data.

1 Introduction

Let $\{x_1, \ldots, x_n\}$ be a set of objects endowed with a discrimination function $\psi(x_i, x_j)$. The primary meaning of $\psi(x_i, x_j)$ in Fechnerian scaling (FS) is the probability with which x_i is judged to be different from (not the same as) x_j. For example, a pair of colors (x_i, x_j) may be repeatedly presented to an observer, and $\psi(x_i, x_j)$ may be estimated by the frequency of responses "they are different." Or, (x_i, x_j) may be a pair of statistical models, and $\psi(x_i, x_j)$ the probability with which model x_j fails to fit (by some statistical criterion) a randomly chosen data set generated by model x_i. Moreover, ψ need not be a probability, it can be any pairwise measure which can be interpreted as the degree of "how dissimilar x_i and x_j are."

FS is aimed at imposing a metric on $\{x_1, \ldots, x_n\}$ based on ψ. Unlike multidimensional scaling (e.g., Kruskal and Wish 1978), FS does not require the ψ-data to satisfy such properties as constant self-dissimilarity ($\psi(x_i, x_i) \equiv \text{const}$) or symmetry ($\psi(x_i, x_j) = \psi(x_j, x_i)$). It is a well-established empirical fact that $\psi(x_i, x_j)$ is not a metric. Discrimination probabilities based on same-different judgments (see the Morse code data used in this paper) systematically violate the

T. Kiefer (✉)
Technische Universität Dortmund, Fakultät Statistik, D-44221 Dortmund, Germany
e-mail: kiefer@statistik.tu-dortmund.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_34,

probability-distance hypothesis, a hypothesis the data-analytic techniques such as multidimensional scaling are based on: for some "true" distance $d(x_i, x_j)$ and some monotone transformation f, $\psi(x_i, x_j) = f(d(x_i, x_j))$. The only property of the ψ-data which is required by FS is regular minimality. In addition to the Fechnerian distances, FS also computes geodesic chains and loops, points of subjective equality, and the generalized Shepardian dissimilarity index. These concepts of FS are explained in Sect. 2.

This paper presents the R (http://www.R-project.org/, R Development Core Team 2009) package **fechner** for FS of object (or stimulus) sets. The package is available on CRAN (http://CRAN.R-project.org/package=fechner). It has functions for checking the required data format and the fundamental regular minimality (or maximality) property. The main function of the package computes the overall Fechnerian distances, and such additional information as the geodesic chains and loops, oriented Fechnerian distances, points of subjective equality, and the generalized Shepardian dissimilarity index. There are plot and summary methods for graphing and outlining the results obtained from FS analyses. Currently available software for FS includes **FSCAMDS**, which runs on MATLAB, and a MATLAB toolbox. This software can be downloaded from http://www.psych.purdue.edu/~ehtibar/ and http://www.psychologie.uni-oldenburg.de/stefan.rach/, respectively.

The paper is structured as follows. In Sect. 2, we briefly review the theory of FS. In Sect. 3, we describe the functions of the package **fechner**. In Sect. 4, we apply the functions to real Morse code data of discrimination probabilities.

2 Theory of FS

For details about the theory of FS, see Dzhafarov and Colonius (2006). The only property of the ψ-data which is required by FS is regular minimality (RM). This property can be formulated in three statements:

(a) For every x_i there is one and only one x_j such that $\psi(x_i, x_j) < \psi(x_i, x_k)$ for all $k \neq j$ (this x_j is called the point of subjective equality, or PSE, of x_i);
(b) For every x_j there is one and only one x_i such that $\psi(x_i, x_j) < \psi(x_k, x_j)$ for all $k \neq i$ (this x_i is called the PSE of x_j);
(c) x_j is the PSE of x_i if and only if x_i is the PSE of x_j.

Every data matrix in which every diagonal entry $\psi(x_i, x_i)$ is smaller than all entries $\psi(x_i, x_k)$ in its row ($k \neq i$) and all entries $\psi(x_k, x_i)$ in its column ($k \neq i$) satisfies RM in the simplest (so-called canonical) form. In this simplest case every object x_i is the PSE of x_i. (Note that regular maximality can be defined analogously, replacing "minimal" with "maximal." This is required when the ψ-data represent closeness values rather than differences: e.g., $\psi(x_i, x_j)$ may be the percent of times x_i is judged to be the same as x_j.)

Given a matrix of $\psi(x_i, x_j)$-values with the rows and columns labeled by $\{x_1, \ldots, x_n\}$, if (and only if) RM is satisfied, the row objects and column

objects can be presented in pairs of PSEs $(x_1, x_{k_1}), (x_2, x_{k_2}), \ldots, (x_n, x_{k_n})$, where $(k_1, k_2, \ldots, k_n)$ is a permutation of $(1, 2, \ldots, n)$. The FS procedure identifies and lists these PSE pairs and then relabels them so that two members of the same pair receive one and the same label:

$$(x_1, x_{k_1}) \mapsto (a_1, a_1), (x_2, x_{k_2}) \mapsto (a_2, a_2), \ldots, (x_n, x_{k_n}) \mapsto (a_n, a_n).$$

For instance, the matrix of ψ-data

$$\begin{bmatrix} & x_1 & x_2 & x_3 \\ x_1 & 0.2 & 0.1 & 0.5 \\ x_2 & 0.7 & 0.3 & 0.2 \\ x_3 & 0.1 & 0.6 & 0.3 \end{bmatrix}$$

becomes

$$\begin{bmatrix} & a_3 & a_1 & a_2 \\ a_1 & 0.2 & 0.1 & 0.5 \\ a_2 & 0.7 & 0.3 & 0.2 \\ a_3 & 0.1 & 0.6 & 0.3 \end{bmatrix} = \begin{bmatrix} & a_1 & a_2 & a_3 \\ a_1 & 0.1 & 0.5 & 0.2 \\ a_2 & 0.3 & 0.2 & 0.7 \\ a_3 & 0.6 & 0.3 & 0.1 \end{bmatrix}$$

in which each diagonal entry is minimal in its row and in its column. After this relabeling the original function $\psi(x_i, x_j)$ is redefined: $p_{ij} = \hat{\psi}(a_i, a_j)$. In the package **fechner** the pairs of PSEs are assigned identical labels leaving intact the labeling of the rows and relabeling the columns with their corresponding PSEs (referred to as canonical relabeling).

FS imposes a metric G on the set $\{a_1, \ldots, a_n\}$ in such a way that, if x_i and $x_{i'}$ are each other's PSEs relabeled into a_i and x_j and $x_{j'}$ are each other's PSEs relabeled into a_j, then $G(x_i, x_j) = G(x_{i'}, x_{j'}) = G(a_i, a_j)$. For every pair of objects (a_i, a_j) we consider all possible chains of objects $(a_i, a_{k_1}, \ldots, a_{k_r}, a_j)$, where $(a_{k_1}, \ldots, a_{k_r})$ is a sequence chosen from $\{a_1, \ldots, a_n\}$. For each such a chain we compute what is called its psychometric length (of the first kind) as

$$L^{(1)}(a_i, a_{k_1}, \ldots, a_{k_r}, a_j) = \sum_{m=0}^{m=r} (p_{k_m k_{m+1}} - p_{k_m k_m}),$$

where we put $a_i = a_{k_0}$ and $a_j = a_{k_{r+1}}$. The quantities $p_{k_m k_{m+1}} - p_{k_m k_m}$ are referred to as psychometric increments of the first kind. Then we find a chain with the minimal value of $L^{(1)}$, and take this minimal value of $L^{(1)}$ for the quasidistance $G_{ij}^{(1)}$ from a_i to a_j (referred to as the oriented Fechnerian distance of the first kind). Quasidistance (quasimetric, or oriented distance) is a pairwise measure which satisfies all metric properties except for symmetry. In FS we symmetrize this quasimetric and transform it into a metric by computing $G_{ij}^{(1)} + G_{ji}^{(1)}$ and taking it for the "true" or "overall" Fechnerian distance (of the first kind) G_{ij} between a_i and a_j. Any chain $(a_i, a_{k_1}, \ldots, a_{k_r}, a_j)$ with $L^{(1)}(a_i, a_{k_1}, \ldots, a_{k_r}, a_j) =$

$G_{ij}^{(1)}$ is called a geodesic chain (of the first kind). Then the overall Fechnerian distance G_{ij} is the psychometric length (of the first kind) of a geodesic loop $(a_i, a_{k_1}, \ldots, a_{k_r}, a_j, a_{l_1}, \ldots, a_{l_s}, a_i)$, or equivalently $(a_j, a_{l_1}, \ldots, a_{l_s}, a_i, a_{k_1}, \ldots, a_{k_r}, a_j)$.

We can also compute the psychometric length (of the second kind) of an arbitrary chain $(a_i, a_{k_1}, \ldots, a_{k_r}, a_j)$ as

$$L^{(2)}(a_i, a_{k_1}, \ldots, a_{k_r}, a_j) = \sum_{m=0}^{m=r} (p_{k_{m+1}k_m} - p_{k_m k_m}),$$

where $p_{k_{m+1}k_m} - p_{k_m k_m}$ are called psychometric increments of the second kind, and then define the quasidistance (the oriented Fechnerian distance of the second kind) $G_{ij}^{(2)}$ from a_i to a_j as the minimal value of $L^{(2)}$ across all chains inserted between a_i and a_j. It makes, however, no difference for the final computation of the overall Fechnerian distance G_{ij} (of the first or second kind), because it can be shown that

$$G_{ij} = G_{ij}^{(1)} + G_{ji}^{(1)} = G_{ij}^{(2)} + G_{ji}^{(2)}.$$

Any geodesic loop $(a_i, a_{k_1}, \ldots, a_{k_r}, a_j, a_{l_1}, \ldots, a_{l_s}, a_i)$ of the first kind, if traversed in the opposite direction, as $(a_i, a_{l_s}, \ldots, a_{l_1}, a_j, a_{k_r}, \ldots, a_{k_1}, a_i)$, becomes a geodesic loop of the second kind.

The package **fechner** compares the value of G_{ij} (referred to as G) to what we call a Shepardian index of dissimilarity S_{ij}: $S_{ij} = p_{ij} + p_{ji} - p_{ii} - p_{jj}$. Note that $G_{ij} \leq S_{ij}$ in all cases. If the geodesic loop for (a_i, a_j) contains no other objects, i.e., if it is (a_i, a_j, a_i), then $G_{ij} = S_{ij}$.

3 The R Package fechner

The package **fechner** is implemented based on the S3 system. It comes with a namespace and consists of three external functions (functions the package exports); they are described below. The package also contains internal functions, which basically are `plot`, `print`, and `summary` methods for objects of the class `fechner`. There are two real and two artificial data sets accompanying this package.

The functions `check.data`

```
check.data(X, format = c("probability.different",
  "percent.same", "general"))
```

and `check.regular`

```
check.regular(X, type = c("probability.different",
  "percent.same", "reg.minimal", "reg.maximal"))
```

are used to check whether the data `X` are of required format and whether `X` satisfy regular minimality/maximality, respectively. The data `X` must be a matrix or data

frame, have the same number of rows and columns, and be numeric (no infinite, undefined, or missing values are allowed). This is the `"general"` data format. The `"probability.different"` and `"percent.same"` formats, in addition, require that the data lie in $[0, 1]$ and $[0, 100]$, and imply checks for regular minimality and regular maximality, respectively. The values `"reg.minimal"` and `"reg.maximal"` can be specified to force checking regular minimality and regular maximality, respectively, independent of the data set used. If all of the requirements for a data format are satisfied, `check.data` returns the data with rows and columns labeled. If the data do satisfy regular minimality/maximality, `check.regular` returns the canonical representation of the data in which regular minimality/maximality is satisfied in canonical form, and the pairs of PSEs with their common labels.

The primary function of the package is `fechner`, which provides all the Fechnerian scaling computations:

```
fechner(X, format = c("probability.different",
  "percent.same", "general"), compute.all = FALSE,
  check.computation = FALSE)
```

The argument `format` and the requirements for the data `X` are the same as before (`fechner` calls the function `check.regular`, which in turn calls `check.data`). The default value `FALSE` for `compute.all` corresponds to short computation, which yields the main Fechnerian scaling computations (see Sect. 4). The value `TRUE` corresponds to long computation, which additionally yields intermediate results and also allows for a check of computations if `check.computation` is set `TRUE`. The performed check computes the difference "overall Fechnerian distance of the first kind minus overall Fechnerian distance of the second kind." The function `fechner` returns an object of the class `fechner`, for which `plot`, `print`, and `summary` methods are provided. The `plot` method graphs the results obtained in the FS analyses. It produces a scatterplot of the overall Fechnerian distance G versus the S-index. The `print` method prints the main results obtained in the FS analyses, which are the overall Fechnerian distances and the geodesic loops. The `summary` method outlines the results obtained in the FS analyses. It returns a list consisting of the pairs of objects and their corresponding S-index and G values, the value of the Pearson correlation coefficient between them, the value of the C-index (as an ad hoc measure of the "improvement" the psychometric increments need to become metric), and the level of comparison chosen. For details, see Ünlü et al. (2009).

4 Example

We use the Rothkopf (1957) Morse code data of discrimination probabilities among 36 auditory Morse code signals for the letters $A, B, \ldots, Z$ and the digits $0, 1, \ldots, 9$ to demonstrate the functions of the package. Each number in the data frame `morse` gives the percentage of subjects who responded "same" to the row signal followed by the column signal.

The data set `morse` satisfies regular maximality in the canonical form:

```
R> check.regular(morse, type = "percent.same")$check
[1] "regular maximality"
R> check.regular(morse,
R+ type = "percent.same")$in.canonical.form
[1] TRUE
```

For typographic reasons only, in the sequel we consider a small subset of this stimulus set, chosen to form a "self-contained" subspace: a geodesic loop for any two elements of such a subset (computed using the complete data set) is contained entirely within the subset. Note that the results obtained from Fechnerian scaling analyses restricted to self-contained subspaces are the same as the results obtained based on the entire stimulus sets. For instance, a particular self-contained 10-code subspace of the 36 Morse codes consists of the codes for the letter B and the digits $0, 1, 2, 4, 5, \ldots, 9$.

```
R> indices <- which(is.element(names(morse),
R+ c("B", c(0, 1, 2, 4:9))))
R> f.scal.morse <- fechner(morse, format = "percent.same")
R> f.scal.morse$geodesic.loops[indices, indices]
```

```
        B    1     2    4    5      6      7       8      9      0
B       B  B1B   B2B B46B  B5B    B6B  B676B B67876B B6789B   B06B
1     1B1    1   121  141  151    161   1781     181    191    101
2     2B2  212     2  242  252    262    272     282   2192  21092
4    46B4  414   424    4  454   46B4    474    4784    494    404
5     5B5  515   525  545    5   56B5    575     585    595    505
6     6B6  616   626 6B46 6B56      6    676   67876 678976    606
7   76B67 7817   727  747  757    767      7     787   7897 789097
8 876B678  818   828 8478  858  87678    878       8    898   8908
9  9B6789  919  9219  949  959 976789   9789     989      9    909
0    06B0  010 09210  040  050    060 097890    0890    090      0
```

The discrimination probabilities for this part of the `morse` data are:

```
R> (morse.subspace <- morse[indices, indices])
   B  1  2  4  5  6  7  8  9  0
B 84 12 17 40 32 74 43 17  4  4
1  5 84 63  8 10  8 19 32 57 55
2 14 62 89 20  5 14 20 21 16 11
4 19  5 26 89 42 44 32 10  3  3
5 45 14 10 69 90 42 24 10  6  5
6 80 15 14 24 17 88 69 14  5 14
7 33 22 29 15 12 61 85 70 20 13
8 23 42 29 16  9 30 60 89 61 26
9 14 57 39 12  4 11 42 56 91 78
0  3 50 26 11  5 22 17 52 81 94
```

We see that the Morse code discrimination probability data violate constant self-dissimilarity. For example, the Morse code for digit 1 was judged different from

itself by 16% of respondents, but only by 6% for digit 0. Symmetry is violated as well: The digits 4 and 5, for instance, were judged to be different in 58% of cases when 4 was presented first, but in only 31% when 4 was presented second.

The function `fechner` is the main function of the package and provides all the Fechnerian scaling computations. The overall Fechnerian distances are:

```
R> f.scal.subspace.mo <- fechner(morse.subspace,
R+ format = "percent.same", compute.all = FALSE,
R+ check.computation = FALSE)
R> f.scal.subspace.mo$overall.Fechnerian.distances
     B    1    2    4    5    6    7    8    9    0
B 0.00 1.51 1.42 0.97 0.97 0.18 0.61 1.05 1.49 1.60
1 1.51 0.00 0.48 1.60 1.50 1.49 1.27 0.99 0.61 0.73
2 1.42 0.48 0.00 1.32 1.64 1.49 1.25 1.28 1.06 1.21
4 0.97 1.60 1.32 0.00 0.68 0.97 1.27 1.45 1.65 1.69
5 0.97 1.50 1.64 0.68 0.00 1.08 1.39 1.60 1.71 1.74
6 0.18 1.49 1.49 0.97 1.08 0.00 0.43 0.87 1.35 1.46
7 0.61 1.27 1.25 1.27 1.39 0.43 0.00 0.44 0.92 1.18
8 1.05 0.99 1.28 1.45 1.60 0.87 0.44 0.00 0.63 0.83
9 1.49 0.61 1.06 1.65 1.71 1.35 0.92 0.63 0.00 0.26
0 1.60 0.73 1.21 1.69 1.74 1.46 1.18 0.83 0.26 0.00
```

The information provided using short computation, an overview (the output is omitted, for typographic reasons):

```
R> attributes(f.scal.subspace.mo)
```

An overview of the information computed under long computation, which additionally yields intermediate results and allows for a check of computations (the output is omitted, for typographic reasons):

```
R> f.scal.subspace.long.mo <- fechner(morse.subspace,
R+ format = "percent.same", compute.all = TRUE,
R+ check.computation = TRUE)
R> attributes(f.scal.subspace.long.mo)
```

Objects of the class `fechner` can be summarized:

```
R> summary(f.scal.morse)

number of stimuli pairs used for comparison: 630

summary of corresponding S-index values:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.180   1.260   1.520   1.435   1.670   1.850

summary of corresponding Fechnerian distance G values:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
    0.180   1.203   1.490   1.405   1.660   1.850

Pearson correlation: 0.9764753

C-index: 0.002925355

comparison level: 1
```

5 Conclusion

We have introduced the package fechner for Fechnerian scaling of object sets in R. The package has functions for checking the required data format and the regular minimality/maximality property, a fundamental property of discrimination in psychophysics. The main function of the package provides all the Fechnerian scaling computations, in the short and long variants.

By contributing the package **fechner** in R we hope to have established a basis for computational work in this field. The realization of Fechnerian scaling in R may prove valuable in applying current or conventional statistical methods to the theory of Fechnerian scaling. For instance, the determination of confidence regions (e.g., for overall Fechnerian distances) and hypothesis testing (e.g., testing for RM) in Fechnerian scaling are likely to be based on resampling methods. Such an endeavor would involve extensive computer simulation, something R would be ideally suited for.

References

Dzhafarov, E. N., & Colonius, H. (2006). Reconstructing distances among objects from their discriminability. *Psychometrika, 71*, 365–386.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional Scaling*. Beverly Hills: Sage.

R Development Core Team. (2009). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rothkopf, E. Z. (1957). A measure of stimulus similarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology, 53*, 94–101.

Ünlü, A., Kiefer, T., & Dzhafarov, E. N. (2009). Fechnerian scaling in R: The package **fechner**. *Journal of Statistical Software, 31*(6), 1–24.

Asymptotic Behaviour in Symbolic Markov Chains

Monique Noirhomme-Fraiture

Abstract Many random processes depending upon time appear in fact in interval even if the standard modelisation reduces it to the mean. It is the case for daily temperature, daily value of stocks. They are known with minimum and maximum values, during a period of time. Thus, they can be considered as multivalued stochastic processes.

In another paper (Noirhomme-Fraiture and Cuvelier 2007) we have introduced the so-called symbolic stochastic processes and more particularly the symbolic Markov chains. In that case, we suppose that the process has no memory of the past. In this paper, after reminding the basic notions for symbolic Markov chains, we will present some asymptotic behaviour for these processes.

1 Introduction

A stochastic process $\underline{X}_t$ is a random variable which depends on time. In the standard theory, the state space can be discrete or continuous but in the real life X_t can also appear as multi-valued or an interval or even an histogram. For example, let us consider the evolution of a stock value, per day: the stock has several (continuous) values: open, close, mean, maximum. We could consider it as an interval value [min, max]. In meteorology, the temperature is also daily given by a minimum and a maximum value. In another domain, the evolution of the daily audience of a TV channel is given by the percentage of time spent at watching the channel. The audience is thus given by a histogram. In all these cases the variables are symbolic ones.

Many books have been written about stochastic processes (Cox and Miller 1965; Karlin 1966; Beichelt 2006; Gihman and Skorohod 2004), but very few has been done on symbolic stochastic ones. De Carvalho et al. (2004) have studied symbolic linear regression. Prudencio et al. (2004) have studied time series. We have introduced the basic concepts of Symbolic Markov Chain in Noirhomme-Fraiture

M. Noirhomme-Fraiture
University of Namur, Namur, Belgium
e-mail: monique.noirhomme@fundp.ac.be

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_35,

and Cuvelier (2007) and given the Chapman Kolmogorov theorem in all the symbolic cases. In this paper, after reminding the basic definitions and results, we study the asymptotic behaviour of a Markov Chain in discrete time for the multi-valued categorical and interval-valued cases.

2 Symbolic Variables

The symbolic data analysis has been studied by E. Diday and his partners in SODAS and ASSO projects (Bock and Diday 2000; Diday and Noirhomme-Fraiture 2008). Let us remind the definition of a symbolic variable.

Let (Ω, A, P) a probability space. A symbolic variable X is a mapping $\Omega \rightarrow \beta$, where β is more general then in the discrete or continuous state space case. We have in fact the different cases:

1. Single valued variable: $\beta \subseteq \mathbb{R}$
2. Single valued categorical: $\beta \subseteq \mathbb{N}$
3. Multi-valued categorical: Let $Y \subseteq \mathbb{N}$, $\beta = P(Y)$ β is the set of Y subsets
4. Interval: Let $Y \subseteq \mathbb{R}$, $\beta = I(Y)$ β is the set of intervals of Y
5. Modal: Let $Y \subseteq \mathbb{N}$ $\beta = M(Y)$ β is the set of a specified non-negative measure on Y. This measure could be a probability or a weighting.

Let $\underline{X}_t$ a stochastic process. If $\underline{X}_t$ is symbolic, we will speak about *symbolic stochastic process*. We intend to generalise the theory of Markov chains to the symbolic case. In particular, we will consider the results on limit distributions.

3 Markov Chains

When studying a stochastic process $\underline{X}_t$, we make different assumptions in order to simplify the problem.

3.1 The Markov Property

$$\begin{aligned} &\Pr[a < \underline{X}_t \leq b \mid \underline{X}_{t_1} = x_1, \underline{X}_{t_2} = x_2, \ldots, \underline{X}_{t_n} = x_n] \\ &\quad = \Pr[a < \underline{X}_t \leq b \mid \underline{X}_{t_n} = x_n],\ t_1 < t_2 < \ldots < t_n < t \end{aligned} \quad (1)$$

express that a Markov chain is a stochastic process without memory of its story. We consider usually that this process is *homogeneous in time*. It means that the transition probabilities $\Pr[\underline{X}_t \in A \mid \underline{X}_s = x]$ are function to $f - s$ and not of s. This property has not to be confused with a *Stationary process* property. For such a process, the joint distribution is the same with a translation in the time

$$\Pr[\underline{X}_{t_1+h} = x_1, \ldots, \underline{X}_{t_n+h} = x_n] = \Pr[\underline{X}_{t_1} = x_1, \ldots, \underline{X}_{t_n} = x_n]. \quad (2)$$

The aim of the theory of stochastic processes in discrete time is to obtain $\Pr[\underline{X}_t \in A \mid \underline{X}_{t-1} = x]$ from the knowledge of the value of $\Pr[\underline{X}_t \in A]$ or to obtain at least asymptotic properties when t tends to ∞ (stationary process case). Let the time *discrete* ($n = 1, 2, \dots$). Let $\underline{X}_n$ a *categorical variable* (categories $j = 1, \dots, k$) at time n. Let $P_{ij}(n) = \Pr[\underline{X}_{m+n} = j \mid \underline{X}_m = i]$ the stationary transition probabilities. They verify the relation

$$P(m+n) = \sum_k P_{ik}(m)\ P_{kj}(n). \tag{3}$$

It is called the *Chapman-Kolmogorov property(CK property)*. Let $P(n) = \{P_{ij}(n)\}$. The Chapman-Kolmogorov property leads to $P(n) = P^n$. Thus from P_{ij} we can compute $P_{ij}(n)$. These properties can be directly applied to *single valued categorical symbolic variables*.

4 The CK Property for Symbolic Stochastic Processes

4.1 Multivalued Categorical Variable

Let $\mathbf{j} = (j_1, \dots, j_s)$ where $j_k = \begin{cases} 1 & \text{if category } C_k \text{ is present,} \\ 0 & \text{elsewhere.} \end{cases}$

The Chapman-Kolmogorov property can be generalised at s-vector states

$$P_{\mathbf{ij}}(n+m) = \sum_{\mathbf{k}} P_{\mathbf{ik}}(n)\ P_{\mathbf{kj}}(m) \tag{4}$$

which allows to compute $P_{\mathbf{ij}}(n)$, knowing $P_{\mathbf{ij}}$.

4.2 Interval Valued Variable

In Noirhomme-Fraiture and Cuvelier (2007) we have considered a single valued variable known in an interval. Here, we modelise an interval-valued variable as a two dimensional variable described by the min and the max of the *random* interval $[\underline{x}_{\min}, \underline{x}_{\max}]$ with the constraint $\underline{x}_{\min} \leq \underline{x}_{\max}$. It is more convenient to use the center $\underline{c}$ and the half length $\underline{h}$ of the interval which has the easier constraint $\underline{h} \geq 0$. Let us consider the two-dimensional stochastic process $\underline{\mathbf{X}} = (\underline{c}_n, \underline{h}_n)$ at time n. Let $\mathbf{x} = (c, h)$, $\mathbf{x}_1 = (c_1, h_1), \dots, \mathbf{x}_n = (c_n, h_n)$. Then $\underline{\mathbf{X}}_n$ is an **interval Markov process** if

$$\begin{aligned} \Pr[\underline{c}_n \leq c, \underline{h}_n \leq h \mid \underline{\mathbf{X}}_1 = \mathbf{x}_1, \dots, \underline{\mathbf{X}}_{n-1} = \mathbf{x}_{n-1}] \\ = \Pr[c_n \leq c, h_n \leq h \mid \underline{\mathbf{X}}_{n-1} = \mathbf{x}_{n-1}]. \end{aligned} \tag{5}$$

Let

$$F((c_0,h_0);(c,h),n) = \Pr[\underline{c}_n \le c, \underline{h}_n \le h \mid \underline{c}_0 = c_0, \underline{h}_0 = h_0] = F(\mathbf{x}_0;\mathbf{x},n), \tag{6}$$

the two dimensional joint conditional distribution for center and half-length. We still suppose that it is homogeneous in time. We can show that

$$F(\mathbf{x}_0;\mathbf{x},m+n) = \int_{-\infty}^{+\infty}\int_{0}^{\infty} d_{c,h}\, F(\mathbf{x}_0;\mathbf{z},m)\, F(\mathbf{z};\mathbf{x},n) \tag{7}$$

with $\mathbf{z} = (c,h)$. The relation (7) is the CK equation for two dimensional continuous variable. Except for very particular cases, these equations can be solved only numerically. **If we consider the particular case where $\underline{\mathbf{c}}_\mathbf{n}$ and $\underline{\mathbf{h}}_\mathbf{n}$ are independent**,

$$\Pr[\underline{c}_n \le c, \underline{h}_n \le h \mid \underline{c}_0 = c_0, \underline{h}_0 = h_0] = \Pr[\underline{c}_n \le c \mid \underline{c}_0]\ \Pr[\underline{h}_n \le h \mid \underline{h}_0]. \tag{8}$$

In this case, the interval-valued process is a juxtaposition of two independent one-dimensional Markov processes and can be studied separately.

5 Stationary Distribution in Discrete Time

For a stationary process, we have a stationary or equilibrium distribution when $\lim_{n\to\infty} P_{i,j}(n)$ exists and is still a probability distribution. Let us see the existence of such a distribution for the different cases.

5.1 Single Categorical Variable

Let us recall some definitions in Markov chain theory:

A class:	the set of states which communicates, it means that there exists n and m such that $P_{i,j}(n) > 0$ and $P_{j,i}(m) > 0$
Recurrent state:	j is recurrent if $\sum_n P_{i,j}(n) = \infty$
Recurrent non-null:	if the mean recurrence time is finite
Ergodic:	aperiodic, recurrent, non-null state

Theorem 1. *For an aperiodic Markov chain with only one final class, the limit $v_j = \lim_{n\to\infty} P_{ij}(n)$ exists and is independent of the initial state. If moreover the class is recurrent non-null, the limit distribution is solution of $\sum v_j = 1$, $v_j = \sum_k v_k$*

P_{kj}, $j \in \beta$ (Feller 1971, p. 393 – Cox and Miller 1965, p. 108). *We can use this result for* ***categorical variables.***

5.2 Multivalued Categorical Variable

When coding the states as explained earlier, we can extend the result to the multivariate case. For an ergodic chain, we have the stationary distribution

$$v_{\mathbf{j}} = \lim_{n\to\infty} P_{\mathbf{i},\mathbf{j}}(n) \tag{9}$$

solution of $\sum v_{\mathbf{j}} = 1,\ v_{\mathbf{j}} = \sum_{\mathbf{k}} v_{\mathbf{k}}\ P_{\mathbf{kj}}$.

5.3 Single Valued Quantitative Variable (Continuous Variable)

Let $p_n(x; y)$ the probability density function at time n

$$p_n(x; y)\ dy = \Pr[y < \underline{X}_n \leq y + dy \mid \underline{X}_0 = x]. \tag{10}$$

The Chapman-Kolmogorov relation can be used for the one step transition probability

$$p_n(x; y) = \int_{-\infty}^{+\infty} p_{n-1}(x; z)\ p(z, y)\ dz. \tag{11}$$

If the equilibrium (or stationary) distribution exists $\lim_{n\to\infty} p_n(x; y) = f(y)$, then it is solution of

$$f(y) = \int_{-\infty}^{+\infty} f(z)\ p(z, y)\ dz. \tag{12}$$

In some particular cases, it may be possible to solve such an equation analytically otherwise one may need to use numerical methods.

5.4 Particular Case: The Random Walk

Let

$$\underline{X}_n = \underline{X}_{n-1} + \underline{Z}_n \tag{13}$$

and let us suppose that the distribution of the increment $\underline{Z}_n$ in one step is only function of the difference between $\underline{X}_n$ and $\underline{X}_{n-1}$. Let $p(z, y) = g(y - z)$. Let μ the mean and σ^2 the variance of $\underline{Z}_n$. We have $\underline{X}_n = \underline{X}_0 + \underline{Z}_1 + \underline{Z}_2 + \ldots + \underline{Z}_n$. X_n is the sum of n independant identically distributed variables and if n is large, X_n

is normally distributed with mean $n\mu$ and variance $n\sigma^2$, by the normal law of large numbers. It can be shown that

- if $\mu > 0$, $\underline{X}_n \to +\infty$, $\lim_{n\to\infty} \Pr[\underline{X}_n > c] = 1$, for all $c > 0$
- if $\mu < 0$, $\underline{X}_n \to -\infty$, $\lim_{n\to\infty} \Pr[\underline{X}_n > -c] = 1$, for all $c > 0$
- if $\mu = 0$, the random walk will be within a distance from its starting point after n jumps. This does not exclude the possibility of large excursions from the origin.

(Cox and Miller, p. 48).

5.5 Interval Valued Variable

We have seen that an interval variable can be considered as a two dimensional Markov process $\underline{\mathbf{X}}_n = (\underline{c}_n, \underline{h}_n)$. For the partial derivatives of F, if they exist, we have

$$p_{m+n}(\mathbf{x}_0; \mathbf{x}) = \int_{-\infty}^{+\infty} \int_0^{\infty} p_m(\mathbf{x}_0; \mathbf{z})\ p_n(\mathbf{z}; \mathbf{x})\ dc\ dh, \tag{14}$$

where $\mathbf{z} = (c, h)$. In particular

$$p_n(\mathbf{x}_0; \mathbf{x}) = \int_{-\infty}^{+\infty} \int_0^{\infty} p_{n-1}(\mathbf{x}_0; \mathbf{z})\ p(\mathbf{z}; \mathbf{x})\ dc\ dh. \tag{15}$$

If the limit $\lim_{n\to\infty} p_n(\mathbf{x}_0; \mathbf{x}) = f(\mathbf{x})$ exists then it is solution of

$$f(\mathbf{x}) = \int_{-\infty}^{+\infty} \int_0^{\infty} f(\mathbf{z})\ p(\mathbf{z}; \mathbf{x})\ dc\ dh. \tag{16}$$

5.6 Particular Case

Let us suppose that $\underline{c}$ and $\underline{h}$ are **independent**

$$p_n(\mathbf{x}_0 : \mathbf{x}) = p_n(c_0; c)\ p_n(h_0; h). \tag{17}$$

Thus, when n tends to ∞, the limit distribution is given by

$$f(\mathbf{x}) = f_c(c)\ f_h(h) \tag{18}$$

with f_c and f_h respectively solutions of the following equations

$$f_c(c) = \int_{-\infty}^{+\infty} f_c(z_1)\ p_c(z_1; c)\ dz_1,\ \ c \in [-\infty, +\infty], \tag{19}$$

$$f_h(h) = \int_0^{\infty} f_h(z_2)\ p_h(z_2; h)\ dz_2,\ \ h \in [0, +\infty]. \tag{20}$$

5.7 Random Walk in Discrete Time

For the **centre of the interval**, we can use the result of the continuous variable case in paragraph 5.3. The case of the **half length** is more complex because it has a reflecting barrier at 0. Let $\underline{h}_n = \underline{h}_{n-1} + Z_n \geq 0$. As said before, for a random walk, $\underline{Z}_n$ are independent, identically distributed of density g, but here $\underline{Z}_n$ are positive. Let

$$\begin{cases} H_n(x) = \Pr[\underline{h}_n \leq x], & x > 0, \\ 0, & x < 0, \\ H_n(0) = \Pr[\underline{h}_n = 0] \ (\neq 0\,), & (*) \end{cases} \tag{21}$$

(*) the distribution has a discontinuity jump at 0. From CK relation

$$H_n(x) = \int_0^\infty H_{n-1}(y)\, g(x-y)\, dy. \tag{22}$$

If $H_n(x) \to H(x)$, $H(x)$ is solution of

$$H(x) = \int_0^\infty H(y)\, g(x-y)\, dy, \quad x \geq 0, \tag{23}$$

$$H(x) = 0, \qquad x \leq 0. \tag{24}$$

It has been shown that there is a unique probability distribution $H(x)$ if $E\ \underline{Z}_n < 0$ and no solution if $E\underline{Z}_n \geq 0$ (Cox and Miller, p. 63). This means that the half length is stable if the increasing process has a negative mean (=negative drift). On the other hand, if $E\underline{Z}_n \geq 0$, then $\underline{h}_n$ will not have an equilibrium distribution as $n \to \infty$.

5.8 Non Independent Random Walk

If c_n and h_n are non independent, but random walk, we can still write

$$\mathbf{X}_n = \mathbf{X}_0 + \mathbf{Z}_1 + \cdots + \mathbf{Z}_n \tag{25}$$

where $\mathbf{Z}_i$ are independent two-dimensional random vectors with a given bivariate distribution. It means

$$(c_n, h_n) = (c_0, h_0) + (\Delta c_1, \Delta h_1) + (\Delta c_2, \Delta h_2) + \ldots \tag{26}$$

with $\mathbf{z}_n = (\Delta c_n, \Delta h_n)$ with mean (μ_1, μ_2) and dispersion matrix Σ

$$\Sigma = \begin{vmatrix} \operatorname{var} \Delta c & \operatorname{covar}(\Delta c, \Delta h) \\ \operatorname{covar}(\Delta c, \Delta h) & \operatorname{var} \Delta h \end{vmatrix}. \tag{27}$$

By the central limit theorem, $\mathbf{X}_n$ is asymptotically distributed as a bivariate normal variable with mean $n\mu$ and dispersion matrix $n\Delta$.

6 Conclusion

We have considered symbolic variables depending on time with a Markovian behaviour. For the symbolic cases, in particular the interval one, we have shown how to get the stationary or equilibrium distribution $\lim_{n\to\infty} P_{ij}(n)$ or $\lim_{n\to\infty} P_n(\mathbf{x}_0, \mathbf{x})$. An explicit solution is given when the center and the half length of the interval are independent. We have also studied the random walk case with independent or dependent variables.

6.1 Future Work

Among others, we intend to present the results for the continuous time. We have also to consider the general case where the centre c and the half-length h are dependent. We will apply the theory to real data.

References

Beichelt, F. (2006). *Stochastic processes in science, engineering and finance*. London: Chapman & Hall.

Bock, H. H., & Diday, E. (2000). *Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data*. Berlin, Heidelberg, New York: Springer Verlag.

Cox, D. R., & Miller, H. D. (1965). *The theory of stochastic processes*. London: Methuen & Co.

De Carvalho, F. A. T., Neto Eufrazi De A. L., & Tenero, C. P. (2004). A new method to fit a linear regression model for interval-valued data. In T. Fruckrirth, S. Brundo and G. Palm (Eds.), *Advances in artificial intelligence: Proceedings of the Twenty Seventh German Conference on Artificial Intelligence* (pp. 295–306). Berlin: Springer-Verlag.

Diday, E., & Noirhomme-Fraiture, M. (2008). *Symbolic data analysis and the sodas software*. Berlin, Heidelberg, New York: Wiley.

Feller, W. (1971). *An introduction to probability theory and its applications*. London: Wiley.

Gihman, I. I., & Skorohod, A. V. (2004). The theory of stochastic process I. *Die grundleheren der mathematischen wissenschaften ineinzeldarstellungen*. Berlin: Springer.

Karlin, S. (1966). *A first course in stochastic processes*. New-York: Academic Press.

Noirhomme-Fraiture, M., & Cuvelier, E. (2007). *Contributions in classification and data analysis* (pp. 103–111), chapter Symbolic Markov Chains. Berlin, Heidelberg, New York: Springer.

Prudencio, R. B. C., Ludernir, T., & De Carvalho, F. A. T. (2004). A model symbolic classifier for selecting time series models. *Pattern Recognition Letters, 25*, 911–921.

An Interactive Graphical System for Visualizing Data Quality–Tableplot Graphics

Waqas Ahmed Malik, Antony Unwin, and Alexander Gribov

Abstract Poor quality, inaccurate or inadequate data can lead to inappropriate assumptions, misleading results, bias and ultimately poor policy and decision making. Finding errors and cleaning data is a time consuming process and requires domain knowledge. This work presents a modified technique – called Interactive Tableplot – for visualizing data and supporting the incorporation of user's domain knowledge so that erroneous cases are easily revealed. Tableplot is implemented in the software Gauguin (Gribov et al. 2006) that provides techniques for the interactive visual exploration of multivariate datasets.

1 Introduction

The increased use of data to inform policy and improve practice requires a renewed emphasis on assuring the underlying accuracy and reliability of data. High quality data are critical for decision making, priority setting, and ongoing monitoring of programs and policies.

Many techniques employed for detecting erroneous data are fundamentally identical but with different names in different fields. For example, in the field of statistics, the term outlier detection has been used for many years to detect and, where appropriate, remove anomalous observations from data. However, in computer sciences or data mining, novelty detection, anomaly detection, noise detection, deviation detection or exception mining are also found. According to Barnett and Lewis (1994) observations which appear to be inconsistent with the rest of the dataset are outliers or potential anomalies.

Graphics have been widely used in detection of outliers. Unfortunately graphical techniques are unable to incorporate more than a few dimensions of different types.

W.A. Malik
Department of Computer Oriented Statistics and Data Analysis, Institute of Mathematics, University of Augsburg, Germany
e-mail: malik@math.uni-augsburg.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_36,

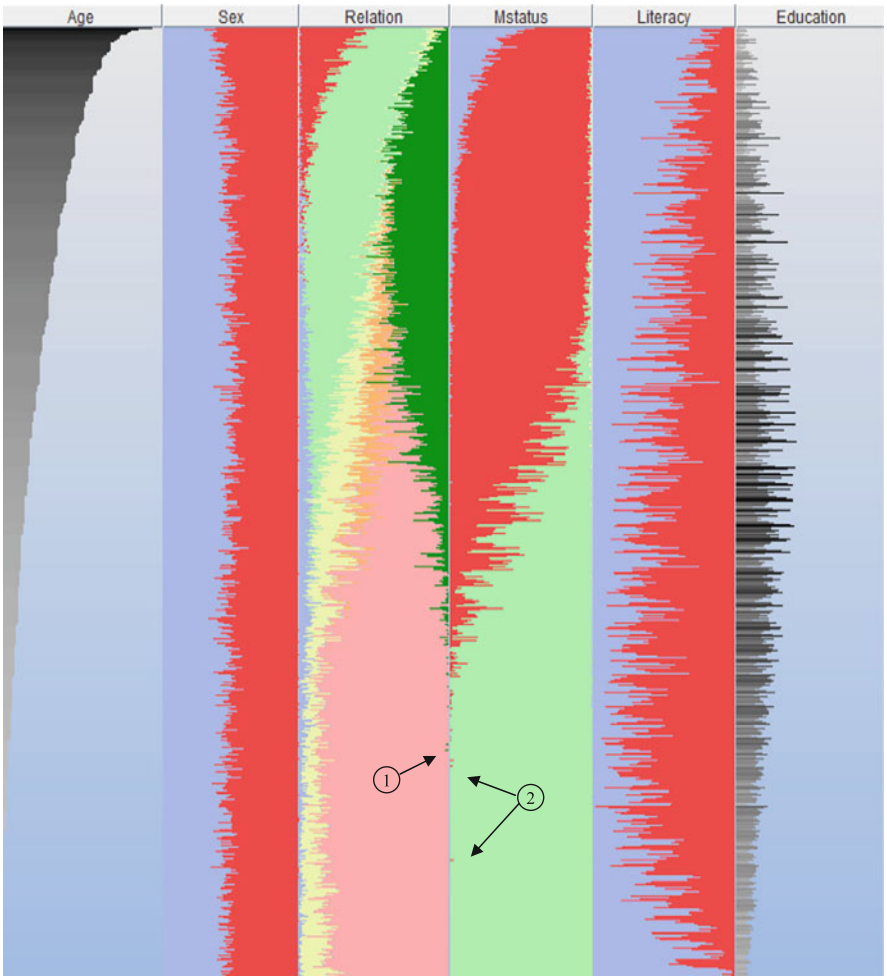


Fig. 1 A Tableplot of the Labour Force Survey Dataset for the demographic variables. The data is sorted by *age* in descending order

To make best use of domain knowledge, the user has to be able to work with different data types and many dimensions simultaneously; and be able to visualize data in such a way that anomalies are emphasized.

Figure 1 shows a Tableplot for the demographic variables of the Labour Force Survey dataset. In this plot, the data are presented in a table-like layout. Each column represents a variable, and each row is an aggregation of 160 cases. The data have been sorted according to *age* in descending order. Querying the graphic shows that the proportion of literates is low for old persons in the dataset. Almost all other variables in the plot show a strong interaction with *age*, as for instance *marital status* and *relation to head* do.

The arrows 1 and 2 show some young persons whose *relation to head* and *marital status* are not consistent with each other. Some very young persons are reported spouse in variable *relation to head* whereas they are "never married" in variable *marital status*. Also there are a few people who are reported "married" in *marital status* while their *relation to head* is "son/daughter (unmarried)". One of these variables contains erroneous data. *Age* and *marital status* are consistent with each other. Therefore it is most likely that relation to head is an error. The Tableplot in Fig. 1 was created using the software Gauguin.

2 Visualization Design

The use of Tableplot not only provides a scale advantage – since the bars can be scaled to one pixel wide without perturbing relative comparisons – but also an exploration advantage, since large numbers of tiny bars can be scanned much more quickly than a bunch of textually represented numbers. In a Tableplot representation each cell in the dataset is mapped to a bar in the visualization with the length and colour of the bar encoding the value of the cell. For continuous variables the length of the bar is proportional to the relative size of the represented value within the variable. Cells containing categorical information are mapped to fixed-length rows whose colour encodes the attribute value. This representation was described in Rao and Card (1994).

Figure 2 shows a schematic representation. Each column in the Tableplot represents a column in the spreadsheet and each row represents a case. When the dataset is large and it is not possible to draw all cases individually on screen, then cases are

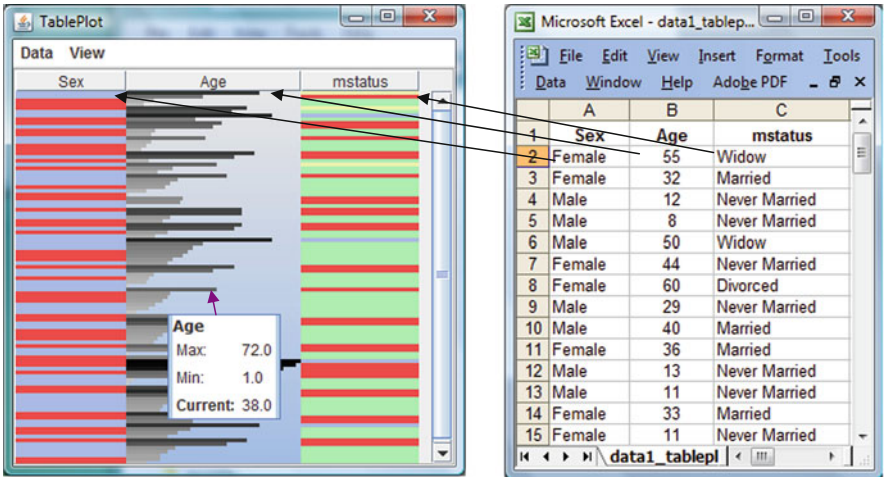


Fig. 2 Visualization design of Tableplot

aggregated. For an aggregated continuous variable, the average value is shown by a horizontal bar of corresponding size (cf. variable *age* in Fig. 1). For an aggregated categorical variable, a horizontal bar is shown, which is subdivided according to the proportion of cases within the particular aggregation group (cf. variable *sex* in Fig. 1).

3 Interactivity in Tableplot

Interactivity is an essential feature of graphics for data exploration, especially in high dimensions. Interactivity features allow using the user's domain knowledge effectively. The following features are implemented in Tableplot.

Sorting: Sorting is an important tool in EDA and is a first step for looking for associations among variables. The data in Tableplot in Fig. 1 have been sorted by *age*, and association of all other variables with *age* is shown, for instance *marital status*, *relation to head* and *literacy*. We can see that marital status is associated with age, and young people are single. Literacy is also associated with age: the level of literacy amongst older people is less than amongst young people.

Zooming: The Tableplot can scale to more rows than vertical pixels available by mapping multiple values into a single line. This is achieved by aggregating neighbouring values. Erroneous cases are usually few in numbers and zooming is important for finding them. Zooming of graphics is achieved by adjusting the row width and aggregation count value.

Querying: Providing different levels of querying is an elegant way of aiding the analyst in an unobtrusive manner (Unwin et al. 2006). Querying a case in Tableplot provides both the value of the selected case and summary statistics of that variable (depending upon the variable type; cf. Fig. 2). In current version of software the querying of continuous variables gives minimum and maximum values but including more statistics like mean, standard deviation and quartiles would help in identifying outliers. Querying allows users to identify outlier values. This helps in deciding what kind of outlier a case is.

4 Tableplot for Visualizing Data Quality

Sorting a column provides powerful support for exploring data. Many properties of the values in a sorted column are apparent by examining the graphical marks (e.g., colour bars) and the shape of the curve in the column. Thus, looking for correlated variables is a matter of scanning across the columns to identify other columns that exhibit a similarly shaped descending curve or one that approximates a mirror image of the curve. This phenomenon helps to find anomalous cases. The cases which don't follow the pattern will be revealed easily.

The Labour Force Survey (LFS) dataset of Pakistan for 2003–04 has been taken to show the power of Tableplot in investigating data quality. Figure 3 shows a Tableplot for seven variables of the LFS dataset for 3,596 regular government employees. The columns include categorical variables *sex, marital status, kind of enterprise* and also numeric variables *age, monthly income, educational level* and *occupation type*. Each row is an aggregation of four cases. Continuous variables are represented by graphical bars proportional in length to the represented values, and categorical values with a corresponding colour. The rows in Fig. 3 are sorted by *age* within *education*. In the dataset occupations are coded from higher to lower i.e. manager is coded as 1 and labour is coded as 33, while level of education is coded from lower to higher i.e., no education is coded as 1 and PhD is coded as 14. The software allows the user to change the position of columns by simple drag and drop. In this way, one can compare particular variables more easily.

In Fig. 3, it is visible that education and occupation are associated with each other. People with high education are working in high level occupations while people with less education are working in low level occupations. In Fig. 3 we can see

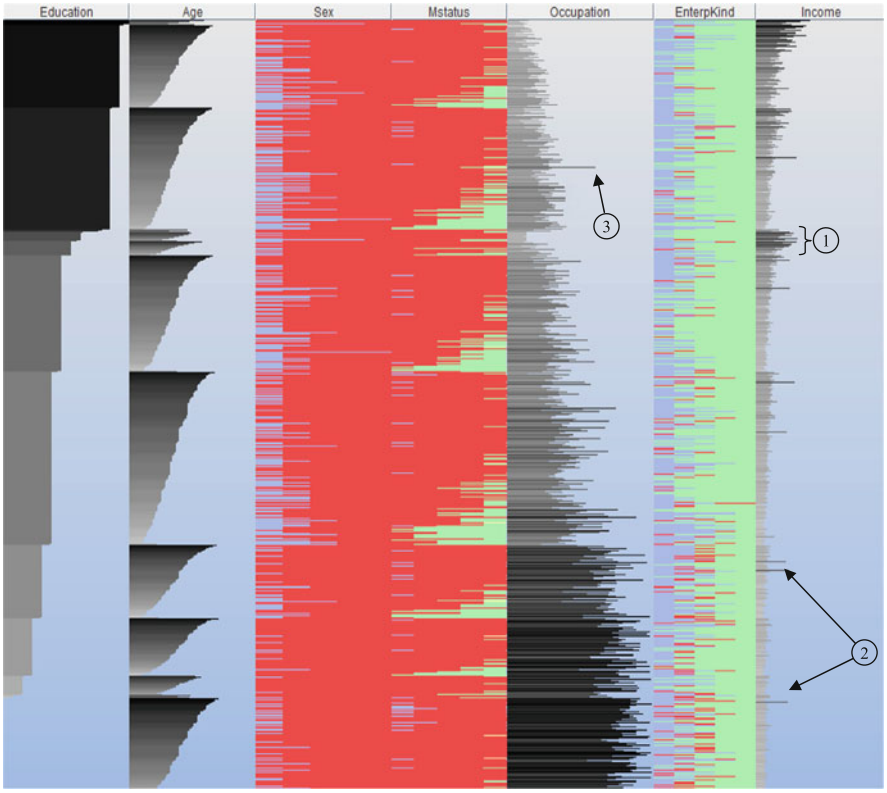


Fig. 3 A Tableplot of the LFS dataset for the employment variables. Rows are sorted by *age* within *education* in descending order

that monthly income declines with occupation level. Arrow 1 shows that persons with medical or engineering degrees have relatively higher monthly income than others. Arrow 2 shows persons with no or low education and working in federal government with very high income. The sudden spikes in the *monthly income* variable are useful for checking consistency of *income* with *education* and *occupation*.

Association between *education* and *occupation* is visible in Fig. 3. Highly educated persons work in managerial and professional occupations while persons with less education work in occupations like industrial worker, machine operator etc. Cases which have no consistency between education, occupation and monthly income are considered to be potentially erroneous cases. Arrow 3 points to highly educated persons who are working in low occupations. Tableplot in Fig. 4 contains the same variables as in Fig. 3 but rows are sorted by *occupation* within *education*.

Arrows 1 and 2 in Fig. 4 show the persons who have no education or very little education but are working in high level occupations like corporate manager or senior legislator, where only a person with high education can work. The income of these

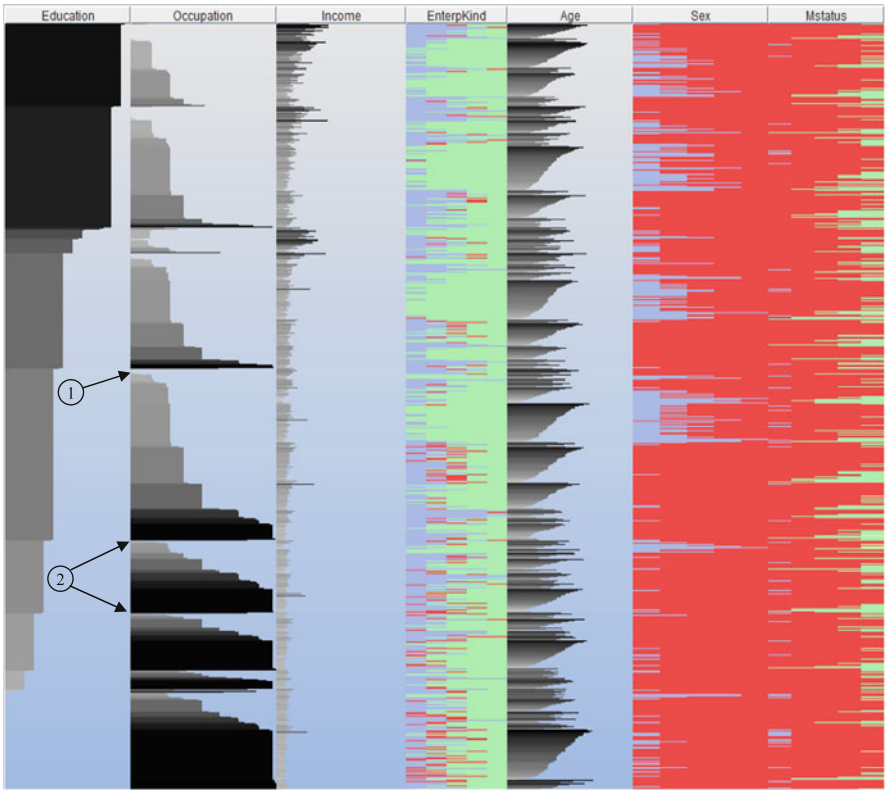


Fig. 4 A Tableplot of same variables as in Fig. 3. Rows are sorted by *occupation* within *education*. Columns are also reordered accordingly

persons is far less than average income of their occupational groups, which means that occupation is in error.

5 Comparison with Other Plots

Other visualization methods for multivariate data include mosaicplots for categorical data and parallel coordinate plots for continuous data. Comparison of Tableplot with these high dimensional plots is performed below.

Mosaicplot: The fluctuation diagram is a variation of mosaicplot, which can be used for visualization of two or more categorical variables (for details see Chen et al. (2008)). Fluctuation diagram of *education* and *occupation* is drawn in Fig. 5. Due to the large difference between the size of biggest and smallest cell, censored zooming is used to zoom in very small cells. In censored zooming, the zooming works by censoring the size of cells at the size of largest cell of the un-zoomed plot [for details see Theus and Urbanek (2008)]. The cells with red borders are censored cells. From this plot we can see that highly educated persons are working in high occupations and low educated persons are working in low occupations. Some peculiar cases are revealed. Persons with low or even no education are working in high occupations (shown by 1), and some highly educated persons are working in low occupations (shown by 2). These small cells which refer to anomalies become visible after censored zooming. Higher dimensional plots allow validating and checking which variable is inconsistent with other variables. Mosaicplots are designed for categorical variables only. Therefore we cannot include continuous variables (e.g. *age* or *monthly income*) in a mosaicplot. However, Tableplot can visualize categorical and continuous variable together.

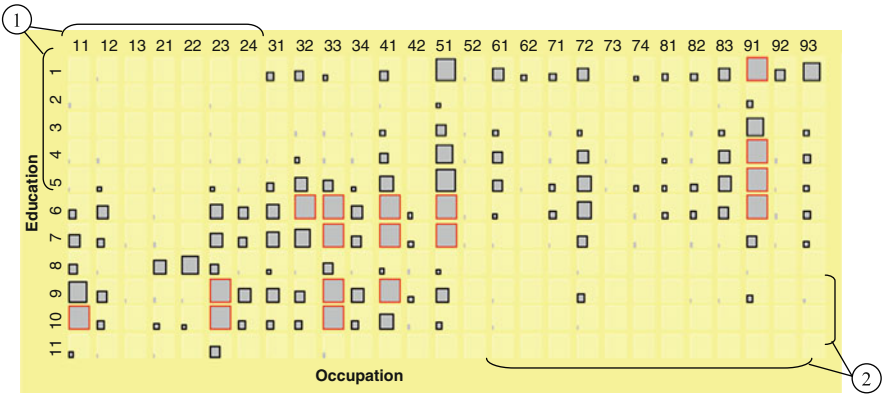


Fig. 5 Fluctuation diagram (a variation of mosaicplot) of *education* and *occupation* from LFS dataset after censored zooming. The plot was drawn with Mondrian Theus (2002)

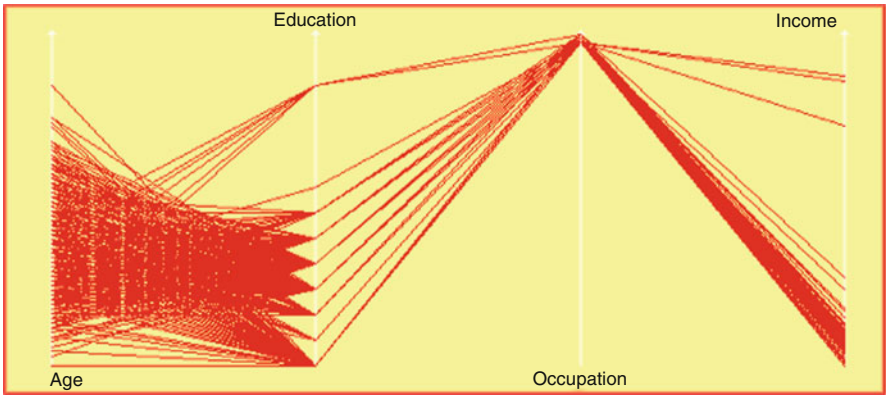


Fig. 6 A Parallel coordinate plot of some variables from LFS dataset. Employees in low level occupations are selected and using hot selection rest of the cases are discarded. The plot was drawn with Mondrian

Parallel Coordinate Plot: Inconsistent cases have been identified by mosaicplot in Fig. 5 but it is not clear which variable contains erroneous values. Introducing other variables like *monthly income* or *monthly household expenditure* will clarify which variable contains peculiar data. Parallel coordinate plot can be used with both continuous and categorical variables. Figure 6 shows a parallel coordinate plot for the variables *age*, *education*, *occupation* and *income*. People working in a low occupation (e.g., machine operator) are selected and using hot selection the remaining cases have been filtered out. The cases which do not have consistent values are shown. Some of the cases with high education also have high income (so probably occupation is wrong). Some of the cases have low income (so education is probably wrong).

6 Software

The Tableplots in this paper were created using the software Gauguin (www.rosuda.org/software/). Gauguin is a standalone package for interactive graphics with an interface to R allowing the seamless integration of some R routines (Urbanek 2003). The major emphasis of the application is to provide interactive graphics for glyphs, and grouping of glyphs by clustering or multi-dimensional scaling. The interactive features include selection, querying, zooming, variation of displays, multiple views and linked highlighting for all plot types. Along with Tableplot, the data can be visualized with different graphics at the same time (e.g., histogram, barchart, scatterplot-matrices, glyphs).

7 Conclusion

Finding anomalies and revealing erroneous data is a difficult and time consuming task. Even a small proportion of bad data can seriously influence analysis results. In many fields, even a very small number of erroneous cases is not acceptable, as in medical and financial data. This paper presents a technique that provides an intuitive visualization approach that could reveal erroneous data more easily. Tableplot along with other interactive plots, available in Gauguin, provides an excellent framework for detecting erroneous data.

Acknowledgements The authors gratefully acknowledge helpful comments by the refrees. This research was funded by Higher Education Commission of Pakistan (HEC) and DAAD.

References

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). New York: Wiley.

Chen, C., Härdle, W., & Unwin, A. (2008). *Handbook of data visualization.* Berlin: Springer-Verlag.

Gribov, A., Unwin, A., & Hofmann, H. (2006). About glyphs and small multiples: Gauguin and the Expo. *Statistical Computing & Statistical Graphics Newsletter, 17*, 18–22.

Rao, R., & Card, S. K. (1994). The table lens: Merging graphical and symbolic representations in an interactive focus+context visualization for tabular information. In ACM SIGCHI *Conference on Human Factors in Computing Systems*.

Theus, M. (2002). Interactive data visualization using Mondrian. *Journal of Statistical Software, 7*, 1–9.

Theus, M., & Urbanek, S. (2008). *Interactive graphics for data analysis: Principles and examples.* London: Chapman & Hall/CRC.

Unwin, A., Theus, M., & Hofmann, H. (2006). *Graphics of large datasets.* New York: Springer.

Urbanek, S. (2003). Rserve-A fast way to provide R functionality to applications. in *Proceedings of 3rd International Workshop on Distributed Statistical Computing*.

Symbolic Multidimensional Scaling Versus Noisy Variables and Outliers

Marcin Pełka

Abstract The aim of the paper is to present and compare effectiveness of symbolic multidimensional scaling methods when we are dealing data with noisy variables and/or outliers. In the article basic terms of symbolic data analysis and symbolic multidimensional scaling are presented.

In empirical part simulation experiment results with application of *Interscal* and *I-Scal* (random and rational start point) are compared based on artificial data (containing noisy variables and/or outliers) generated by `cluster.Gen` procedure from `clusterSim` package of **R** software.

1 Introduction

Symbolic multidimensional scaling aims to present relations between objects treated as hypercubes in multidimensional space. To allow interpretation and graphical representation of the results usually two-dimensional space is used. Most of symbolic multidimensional scaling methods require interval dissimilarity matrix as input. This matrix can be obtained from n judges, opinions or from dissimilarity measure for interval-valued variables that produces interval-valued dissimilarities (see Lechevallier 2001).

In the first part of the paper basic terms of symbolic data, like symbolic variable, symbolic objects, are described. Second part of the article presents four multidimensional scaling methods: *SymScal* and *I-Scal* proposed by Groenen et al. (2005, 2006) and *Interscal* proposed by Denœux and Masson (2000) and an adaptation of Sammon's nonlinear mapping for symbolic objects (Sammon 1969). The third part presents models with noisy variables and/or outliers. In the empirical part results of symbolic multidimensional scaling of artificial data (with noisy variables and/or outliers) are compared.

M. Pełka
Department of Econometrics and Computer Science, Wroclaw University of Economics, Poland
e-mail: marcin.pelka@ue.wroc.pl

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_37,

2 Symbolic Data

Bock and Diday have defined five different symbolic variable types (Bock and Diday 2000, p. 2):

1. single quantitative value,
2. categorical value,
3. quantitative variable of interval type,
4. set of values or categories (multivalued variable),
5. set of values or categories with weights (multivalued variable with weights),
6. modal interval-valued variable proposed in Billard and Diday (2006).

Regardless of their type symbolic variables also can be (Bock and Diday 2000, p. 2):

- taxonomic – which presents prior known structure,
- hierarchically dependent – rules which decide if a variable is applicable or not have been defined,
- logically dependent – logical rules which affect variableŠs values have been defined.

There are two main symbolic objects types:

1. First order objects (simple objects) – single respondent, product, company (single individuals) described by symbolic variables (see Table 1) This objects are individuals that are symbolic by their nature, e.g. a questionnaire might contain such questions as: "please select additional features of a car, that you will buy", "please indicate the interval of a price, that you want to pay for a new car", and so on.
2. Second order objects (aggregate objects, super individuals) – more or less homogeneous classes, groups of individuals described by symbolic variables (see Table 2).

Table 1 Table of symbolic objects (first order objects)

Symbolic objects	Symbolic variables			
	Owned car mark	Preferred car price	Preferred car mark	Preferred colours
Responder 1	Audi	<65000; 80000>	{80% Audi; 20% Toyota}	{blue}
Responder 2	Opel	<30000; 50000>	{60% VW, 30% Audi; 10% Skoda}	{green, white}
Responder 3	Skoda	<28000; 37000>	{60% Honda; 40% Toyota}	{red, white, yellow}
Responder 4	Skoda	<33000; 58000>	{80% Audi, 15% Opel, 5% Toyota}	{black, yellow}
Responder 5	Audi	<66000; 90000>	{65% Audi; 35% VW}	{green}
Responder 6	Opel	<25000; 36000>	{100% Opel}	{black, red}

Source: own research.

Table 2 Table of symbolic objects (second order objects)

Symbolic objects	Symbolic variables		
	Preferred price	Preferred car mark	Preferred colours
Audi	<65000; 90000>	{80% Audi; 20% Toyota; 35% VW}	{blue, green}
Opel	<25000; 50000>	{100% Opel; 60% VW; 30% Audi; 10% Skoda}	{green, white, black, red}
Skoda	<28000; 58000>	{80% Audi; 60% Honda; 40% Toyota; 15% Opel}	{red, white, yellow, black}

Source: own research.

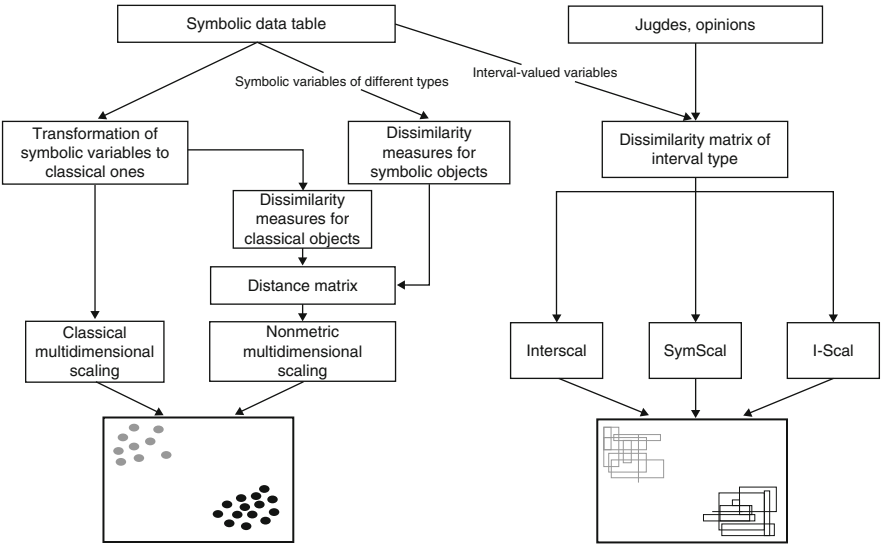


Fig. 1 Symbolic multidimensional scaling methods. *Source:* own research based on Denœux and Masson (2000); Groenen et al. (2005, 2006)

3 Symbolic Multidimensional Scaling Methods

Figure 1 presents main two main approaches ins symbolic multidimensional scaling, classical approach – also known as "symbolique-numerique-symbolique" proposed by E. Diday in 1978 – and symbolic multidimensional scaling based on interval-valued distances. The Fig. 1 presents also main methods for each approach.

The classical approach is based on transformation of symbolic variables to classical ones. It allows to present symbolic objects as points, but transformation causes some information loss about original data structure. Methods based on symbolic dissimilarity measures don't cause loss of information, but they also treat symbolic

objects as points. Symbolic objects shouldn't be treated as points due to the fact that they are not points in multidimensional space. That's why symbolic multidimensional scaling methods based on interval-valued dissimilarities should be applied. In the empirical part Interscal and I-Scal (with random and rational start point) methods were applied. For Interscal I-STRESS loss function was applied. All experiments were done in **R** software with application of source codes written by author.

Algorithm of Interscal method (Denœux and Masson 2000; Lechevallier 2001):

1. Obtain interval-valued dissimilarities, from interval-valued variables or judgements, opinions of n respondents, experts, etc.
2. Construct Δ matrix of interval-valued dissimilarities, where $\overline{\delta}_{ij}$ means lower bound of dissimilarity between ith and jth object; $\underline{\delta}_{ij}$ means upper bound of dissimilarity between ith and jth object.
3. Construct $\widetilde{\Delta}$ matrix defined as follows:

$$\widetilde{\Delta} = \begin{bmatrix} 0 & 0 & \underline{\delta}_{12} & \frac{\overline{\delta}_{12}+\underline{\delta}_{12}}{2} & \underline{\delta}_{13} & \frac{\overline{\delta}_{13}+\underline{\delta}_{13}}{2} & \cdots & \underline{\delta}_{1n} & \frac{\overline{\delta}_{1n}+\underline{\delta}_{1n}}{2} \\ 0 & 0 & \frac{\overline{\delta}_{12}+\underline{\delta}_{12}}{2} & \overline{\delta}_{12} & \frac{\overline{\delta}_{13}+\underline{\delta}_{13}}{2} & \overline{\delta}_{13} & \cdots & \frac{\overline{\delta}_{1n}+\underline{\delta}_{1n}}{2} & \overline{\delta}_{1n} \\ \underline{\delta}_{21} & \frac{\overline{\delta}_{21}+\underline{\delta}_{21}}{2} & 0 & 0 & \underline{\delta}_{23} & \frac{\overline{\delta}_{23}+\underline{\delta}_{23}}{2} & \cdots & \underline{\delta}_{2m} & \frac{\overline{\delta}_{2m}+\underline{\delta}_{2m}}{2} \\ \frac{\overline{\delta}_{21}+\underline{\delta}_{21}}{2} & \overline{\delta}_{21} & 0 & 0 & \frac{\overline{\delta}_{23}+\underline{\delta}_{23}}{2} & \overline{\delta}_{23} & \cdots & \frac{\overline{\delta}_{23}+\underline{\delta}_{23}}{2} & \overline{\delta}_{2m} \\ \underline{\delta}_{31} & \frac{\overline{\delta}_{31}+\underline{\delta}_{31}}{2} & \underline{\delta}_{32} & \frac{\overline{\delta}_{32}+\underline{\delta}_{32}}{2} & 0 & 0 & \cdots & \underline{\delta}_{3m} & \frac{\overline{\delta}_{3m}+\underline{\delta}_{3m}}{2} \\ \frac{\overline{\delta}_{31}+\underline{\delta}_{31}}{2} & \overline{\delta}_{31} & \frac{\overline{\delta}_{32}+\underline{\delta}_{32}}{2} & \overline{\delta}_{32} & 0 & 0 & \cdots & \frac{\overline{\delta}_{3m}+\underline{\delta}_{3m}}{2} & \overline{\delta}_{3m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \underline{\delta}_{m1} & \frac{\overline{\delta}_{m1}+\underline{\delta}_{m1}}{2} & \underline{\delta}_{m2} & \frac{\overline{\delta}_{m2}+\underline{\delta}_{m2}}{2} & \underline{\delta}_{m3} & \frac{\overline{\delta}_{m3}+\underline{\delta}_{m3}}{2} & \cdots & 0 & 0 \\ \frac{\overline{\delta}_{m1}+\underline{\delta}_{m1}}{2} & \overline{\delta}_{m1} & \frac{\overline{\delta}_{m2}+\underline{\delta}_{m2}}{2} & \overline{\delta}_{m2} & \frac{\overline{\delta}_{m3}+\underline{\delta}_{m3}}{2} & \overline{\delta}_{m3} & \cdots & 0 & 0 \end{bmatrix} \tag{1}$$

4. Find the matrix $\mathbf{B} = -\frac{1}{2}\mathbf{J}\widetilde{\mathbf{\Delta}}^{(2)}\mathbf{J}$ with $\mathbf{J}$ the centering matrix.
5. Find eigenvalues $\mathbf{\Phi}^2$ and eigenvectors $\mathbf{P}$ of matrix $\mathbf{B}$.
6. Compute $2n$ points in S-dimensions using the formula: $y_{is} = p_{is}\phi_{ss}$ for $i = 1, 2, \ldots, 2n$ and $s = 1, 2, \ldots, S$.
7. When applying I-STRESS loss function construct the center coordinates $\mathbf{X}$ and spreads of rectangle $\mathbf{R}$ for each object i and each dimension s as follows:

$$x_{is} = \frac{(y_{2i,s} + y_{2i+1,s})}{2} \tag{2}$$

$$r_{is} = \frac{|y_{2i,s} - y_{2i+1,s}|}{2} \tag{3}$$

8. Compute I-STRESS.

Algorithm of I-Scal method (Groenen et al. 2006):

1. Obtain interval-valued dissimilarities, from interval-valued variables or judgments, opinions of n respondents, experts, etc.

2. Set matrix $\mathbf{X_0}$ to initial matrix for coordinate centers of rectangles (I-Scal random start point).
   For I-Scal rational start point obtain matrix $\mathbf{X_0}$ from Interscal.
3. Set matrix $\mathbf{R_0}$ to initial matrix of nonnegative values for the rectangle width.
   For I-Scal rational start point obtain matrix $\mathbf{R_0}$ from Interscal.
4. Set maximum iteration number t and the convergence criterion ε to a small positive value e.g. 10^{-6}.
5. Set iteration counter $k = 0$ and $\mathbf{X}_{-1} = \mathbf{X_0}$, $\mathbf{R}_{-1} = \mathbf{R_0}$.
6. While I-STRESS$_{k-1}$ − I-STRESS$_k \leq \varepsilon$ and $k \leq t$:
7. $k = k + 1$
8. Set $\mathbf{Y}_k = \mathbf{X}_{k-1}$ and $\mathbf{Q}_k = \mathbf{R}_{k-1}$.
   For every dimension:
9. Compute $A_s^{(1)}$ and $B_s^{(1)}$ [see: Groenen et al. (2006) for details].
10. Compute and update matrix of coordinate centers for rectangles $\mathbf{X}$.
11. Compute $A_s^{(2)}$ and $b_s^{(2)}$ [see: Groenen et al. (2006) for details].
12. Compute and update matrix of nonnegative values for the rectangle width $\mathbf{R}$.
13. Set $\mathbf{X}_k = \mathbf{X}$ and $\mathbf{R}_k = \mathbf{R}$.

The I-STRESS loss function, which takes values between 0 and 1, is defined as follows (Groenen et al. 2006; Lechevallier 2001):

$$\text{I-STRESS} = \frac{\sigma_I^2(\mathbf{X}, \mathbf{R})}{\sum_{i<j}^{n} w_{ij}[\delta_{ij}^{(U)}]^2 + \sum_{i<j}^{n} w_{ij}[\delta_{ij}^{(L)}]^2} \tag{4}$$

where: $\sigma_I^2(\mathbf{X}, \mathbf{R}) = \sum_{i<j}^{n} w_{ij}[\delta_{ij}^{(U)} - d_{ij}^{(U)}(\mathbf{X}, \mathbf{R})]^2 + \sum_{i<j}^{n} w_{ij}[\delta_{ij}^{(L)} - d_{ij}^{(L)}(\mathbf{X}, \mathbf{R})]^2$;
$\mathbf{X}$, $\mathbf{R}$ – matrix of rectangles centers ($\mathbf{X}$) and matrix of rectangles span ($\mathbf{R}$);
w_{ij} – weights;
$\delta_{ij}^{(U)}$ and $\delta_{ij}^{(L)}$ – upper and lower distances between ith and jth hiperrectangles;
$d_{ij}^{(U)}$ and $d_{ij}^{(L)}$ – upper and lower distances between rectangles.

4 The Models

For empirical part of the paper four models, each containing 100 objects described by interval-valued variables, were generated with application of `cluster.Gen` procedure from `clusterSim` package of **R** software.

Model I – Three elongated clusters in two dimensions. The observations are independently drawn from bivariate normal distribution with means (0, 0), (1.5, 7), (3, 14) and covariance matrix $\sum(\sigma_{jj} = 1, \sigma_{jl} = 0.9)$.

Model II – Three elongated clusters in three dimensions. The observations are independently drawn from multivariate normal distribution with means(1.5, 6, 3), (3, 12, 6), (4.5, 19, 9), and identity covariance matrix, where $\sum \sigma_{jj} = 1, (1 \leq j \leq 3)$, $\sigma_{12} = \sigma_{13} = 0.9$, and $\sigma_{23} = 0.9$.

Model III – Five clusters in two dimensions that are not well separated. The observations are independently drawn from bivariate normal distribution with means (5, 5), (3, 3), (3, 3), (0, 0), (5, 5), and identity covariance matrix $\sum(\sigma_{jj} = 1, \sigma_{jl} = 0.9)$.

Model IV – Five clusters in three dimensions that are not well separated. The observations are independently drawn from multivariate normal distribution with means (5, 5, 5), (3, 3, 3), (3, 3, 3), (0, 0, 0), (5, 5, 5), and covariance matrix $\sum$, where $\sigma_{jj} = 1$ $(1 \leq j \leq 3)$, and $\sigma_{jl} = 0.9$ $(1 \leq j \neq l \leq 3)$.

To obtain symbolic interval data the data were generated for each model twice into sets A and B and minimal (maximal) value of $\{x_{ij}^{A}; x_{ij}^{B}\}$ is treated as the beginning (the end) of an interval. The noisy variables are simulated independently from the uniform distribution. We require that the variations of noisy variables in the generated data are similar to non-noisy variables. Outliers (for metric and symbolic interval data only). The outliers are generated independently for each variable for the whole data set from uniform distribution with range [1, 10]. The generated values are randomly added to maximum of jth variable or subtracted from minimum of jth variable.

Paths of simulations:

- pure model with no noisy variables and/or outliers,
- model with one noisy variable added,
- model with two noisy variables added,
- model with five noisy variables added,
- model with 20% outliers addes,
- model with one noisy variable and 20% of outliers,
- model with two (three) noisy variables and 20% of outliers,
- model with five noisy variables and 20% of outliers.

5 Results of Simulations

For each model and path of simulation (see: Sect. 4) fifty runs of symbolic multidimensional scaling where done and the means and standard deviations of I-STRESS values where compared.

While Table 3 presents results of simulation for models with known number of noisy variables or models with 20% of outliers, Table 4 presents results of simulations for models with 20% of outliers and known number of noisy variables.

Table 3 Results of simulations – models with noisy variables or 20% of outliers added

Model no.	Method	Mean and SD	Number of noisy variables				20% of outliers
			0	1	2	5	
I	Interscal	Mean	0.3391	0.3992	0.3120	0.4298	0.3734
		SD	0.2153	0.1860	0.1335	0.1356	0.2896
	I-Scal[a]	Mean	0.1333	0.3086	0.2461	0.3143	0.2998
		SD	0.0317	0.1534	0.1291	0.1209	0.2092
	I-Scal[b]	Mean	0.0981	0.4426	0.2344	0.2978	0.2714
		SD	0.0300	0.1544	0.1201	0.1154	0.2765
II	Interscal	Mean	0.3553	0.4221	0.3678	0.4001	0.4109
		SD	0.1362	0.1131	0.1452	0.1653	0.1896
	I-Scal[a]	Mean	0.1209	0.2763	0.2577	0.3243	0.3123
		SD	0.0071	0.1222	0.1098	0.1473	0.1720
	I-Scal[b]	Mean	0.0798	0.3663	0.1876	0.2649	0.2808
		SD	0.0300	0.1544	0.1201	0.1154	0.0987
III	Interscal	Mean	0.4239	0.5049	0.4200	0.3654	0.3987
		SD	0.1667	0.1532	0.1211	0.0973	0.2234
	I-Scal[a]	Mean	0.1092	0.1763	0.2097	0.1980	0.2034
		SD	0.0635	0.0976	0.1112	0.1092	0.1331
	I-Scal[b]	Mean	0.0991	0.1378	0.1876	0.1765	0.2008
		SD	0.0900	0.0865	0.1098	0.1021	0.1109
IV	Interscal	Mean	0.4092	0.5112	0.3961	0.3098	0.4409
		SD	0.1441	0.1110	0.1481	0.2022	0.2344
	I-Scal[a]	Mean	0.1949	0.2109	0.1782	0.2031	0.3092
		SD	0.0742	0.1021	0.1073	0.1143	0.1108
	I-Scal[b]	Mean	0.1302	0.1992	0.1444	0.1673	0.2655
		SD	0.01	0.1009	0.1121	0.1	0.0873

SD – standard deviation
[a] I-Scal with random start point;
[b] I-Scal with rational start point, see: Sect. 3
Source: own research

6 Final Remarks

In this paper several symbolic multidimensional scaling methods were compared on artificially generated symbolic data sets. The experiment showed that the most adequate one for this kind of data in most cases is I-Scal with rational star point, sometimes I-Scal with random star point archives better results as I-Scal with rational start point. When considering all models and paths of simulations I-Scal with rational start point gets the best results and Interscal gets always the worst results. It's suggested to use I-Scal (with random and rational start point) in symbolic multidimensional scaling method in case of real symbolic data.

Note that when we deal table of symbolic objects described by different kinds of symbolic variables (intervals, set of categories, etc.) and we want to preserve all information about the objects, the only way is to treat them as points and apply well-known non-metric multidimensional scaling for classical data. But symbolic objects

Table 4 Results of simulations – models with 20% of outliers and number of noisy variables added

Model no.	Method	Mean and SD	Number of noisy variables			
			1	2	3	5
I	Interscal	Mean	0.5954	0.6003	0.6970	0.8393
		SD	0.1884	0.2073	0.1137	0.1205
	I-Scal[a]	Mean	0.3604	0.4889	0.6659	0.8318
		SD	0.1167	0.1143	0.1656	0.0746
	I-Scal[b]	Mean	0.2530	0.5339	0.6567	0.7629
		SD	0.0480	0.2025	0.1733	0.1310
II	Interscal	Mean	0.5402	0.6094	0.7642	0.8875
		SD	0.1986	0.2772	0.2091	0.0987
	I-Scal[a]	Mean	0.4091	0.5004	0.6109	0.6812
		SD	0.0923	0.1556	0.2441	0.1771
	I-Scal[b]	0.3992	Mean	0.5618	0.6099	0.5665
		SD	0.0791	0.1780	0.1029	0.1245
III	Interscal	Mean	0.4112	0.5443	0.6732	0.8092
		SD	0.1552	0.2007	0.1982	0.2567
	I-Scal[a]	Mean	0.3507	0.4172	0.5094	0.6110
		SD	0.1010	0.1002	0.1788	0.2019
	I-Scal[b]	Mean	0.3678	0.4221	0.4785	0.5012
		SD	0.1209	0.1021	0.1102	0.1761
IV	Interscal	Mean	0.4509	0.5588	0.6789	0.8195
		SD	0.1408	0.1062	0.1343	0.1555
	I-Scal[a]	Mean	0.4021	0.5678	0.6201	0.7985
		SD	0.1231	0.1092	0.0861	0.1023
	I-Scal[b]	Mean	0.4056	0.5401	0.5903	0.7109
		SD	0.1019	0.0989	0.1009	0.1102

SD – standard deviation;
[a] I-Scal with random start point;
[b] I-Scal with rational start point, see: Sect. 3
Source: own research.

due to their complexity shouldn't be treated as points in multidimensional space and lower spaces as well.

References

Billard, L., & Diday, E. (Eds.). (2006). *Symbolic data analysis: Conceptual statistics and data mining*. Chichester: Wiley.

Bock, H.-H., & Diday, E. (Eds.). (2000). *Analysis of symbolic data. Explanatory methods for extracting statistical information from complex data*. Berlin-Heidelberg: Springer Verlag.

Denœux, T., & Masson, M. (2000). Multidimensional scaling of interval-valued dissimilarity data. *Pattern Recognition Letters, 21*(1), 83–92.

Groenen, P. J. F., Winsberg, S., Rodríguez, O., & Diday, E. (2005). Multidimensional scaling of interval dissimilarities. *Econometric Report*, 2005–15, Rotterdam: Erasmus University.

Groenen, P. J. F., Winsberg, S., Rodriguez, O., & Diday, E. (2006). I-Scal: multidimensional scaling of interval dissimilarities. *Computational Statistics and Data Analysis, 51*, 360–378.

Lechevallier, Y. (Ed.). *Scientific report for unsupervised classification, validation and cluster representation.* (2001). Analysis system of symbolic official data – Project number IST-2000-25161, project report.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers, C-18*(5), 401–409.

Principal Components Analysis for Trapezoidal Fuzzy Numbers

Alexia Pacheco and Oldemar Rodríguez

Abstract Scientists in many disciplines face the problem of interpretation of complex structures such as the symbolic data extracted from databases with a significant amount of records; or containing fuzzy numbers based on expert knowledge or partial knowledge originating from incomplete records. Principal Components Analysis (PCA) is most often used to interpret complex patterns as it allows reducing dimensionality and extracting the main characteristics of the data sample, as well as visualization in a two-dimensional plane and in a correlation circle. There is a need to extend this widely used method to the above mentioned data types.

A new method called `PCA-TF` is proposed that allows performing PCA on data sets of trapezoidal (or triangular) fuzzy numbers, that may contain also real numbers and intervals. The approach is an extension to fuzzy numbers of the algorithm by Rodríguez (2000). A group of orthogonal axes is found that permits the projection of the maximum variance of a real numbers' matrix, where each number represents a trapezoidal fuzzy number. The initial matrix of fuzzy numbers is projected to these axes by means of fuzzy number arithmetic, which yields Principal Components and they are also fuzzy numbers. Based on these components it is possible to produce graphs of the individuals in the two-dimensional plane. It is also possible to evaluate the shape of the ordered pairs of fuzzy numbers and visualize the membership function for each point on the z axis over the two-dimensional xy plane. The application is demonstrated on a data sample of students' grades in Denœux and Masson (2004) and is compared to the results of the principal component analysis of fuzzy data using associative neural networks proposed by Denœux & Masson (`D&M-PCA`). Also, an important relation between the arithmetics of the intervals and projection formulas for the interval data type is demonstrated.

A. Pacheco (✉)
Costarican Institute of Electricity, San José, Costa Rica
e-mail: apacheco@ice.go.cr

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_38,

1 Introduction

Recent developments in informatics and statistics opened a possibility of assessing data types with more complex structures such as symbolic objects (Bock and Diday 2000) extracted from a huge amount of records or fuzzy numbers (Kaufman and Gupta 1991) generated on the basis of expert knowledge or partial knowledge originating from incomplete records. It is therefore necessary to extend the classical methods of data analyses to these new data types. One of the most frequently used methods is the Principal Component Analysis (PCA) as it allows dimension reduction and visualization of the results in a low-dimensional space and as a correlation circle. In the context of symbolic data analysis and fuzzy data analysis there have been a number of attempts to extend this method to cover the new data types, for example, intervals (Cazes et al. 1997; Rodríguez 2000) and trapezoidal fuzzy numbers based on associative neural networks (Denœux and Masson 2004).

Herein a new approach, `PCA-TF`, to perform PCA on a trapezoidal fuzzy number matrix is suggested. It considers the theoretical aspects of several independent topics that all operate on interval variable types such as fuzzy numbers, interval arithmetic and symbolic data. An important relation between the arithmetic of the intervals and projection formulas for the interval data type used in the context of symbolic data analysis is demonstrated. A relationship between fuzzy numbers and intervals was already noted by Lodwick and Jamison (2003). Two computer programs (Edit-PCA-TF, `PCA-TF`) were developed to apply and to test the method on arbitrary data sets. The application is shown on the example of the data matrix with student grades found in Denœux and Masson (2004).

2 The Method: PCA-TF

As in the case of an usual PCA, the objective of `PCA-TF` is to obtain a low-dimensional representation of the objects/individuals with minimum information loss, which facilitates compression of the initial data and extraction of the most relevant characteristics. The new development is an extension of the algorithm by Rodríguez (2000)(Cazes et al. 1997) to fuzzy numbers. A set of orthogonal axes is found that allows projecting the maximum variance of a real matrix that corresponds to the middle points of the mean intervals representing each trapezoidal fuzzy number in the most natural way, as noted by Dubois (2006). The initial fuzzy numbers matrix is projected on these new axes by means of fuzzy numbers arithmetic, this yields principal components that are fuzzy numbers as well. Based on these components, it is possible to plot the individuals in the principal component plane and also appreciates the shape of the ordered fuzzy number pairs. Besides it is possible to visualize in a 3D space (XYZ) the principal component plane ($xy-axis$) and the value of the membership function on the z-axis.

To extend the algorithm by Rodríguez (2000) (Cazes et al. 1997), first a relationship among the equations for the projection of an interval type variable and

the interval arithmetic was proved, specifically in [Rodríguez 2000, Theorem 4.1.1] where it is established that if an interval type variable which geometrically is a hypercube defined by the i-th column of the matrix Z is projected over the j-th principal component (in the direction u_j, where u_j is a real number vector), then the maximum ($\overline{r_{ij}}$) and minimum ($\underline{r_{ij}}$) values are defined by the following equations:

$$\underline{r_{ij}} = \sum_{k=1,u_{kj}>0}^{p} \underline{Z_{ki}} \cdot u_{kj} + \sum_{k=1,u_{kj}<0}^{p} \overline{Z_{ki}} \cdot u_{kj}, \tag{1}$$

$$\overline{r_{ij}} = \sum_{k=1,u_{kj}>0}^{p} \overline{Z_{ki}} \cdot u_{kj} + \sum_{k=1,u_{kj}<0}^{p} \underline{Z_{ki}} \cdot u_{kj}, \tag{2}$$

where Z_i represents an interval type variable and $Z_{ji} = [\underline{Z}_{ji}\, \overline{Z}_{ji}]$.

Using $X = Z^t$ ($X_{ij} = Z_{ji}$), $y_{ij} = r_{ij}$, $y_{ij} = [\underline{y_{ij}}, \overline{y_{ij}}]$ and $r_{ij} = [\underline{r_{ij}}, \overline{r_{ij}}]$ then equations (1) and (2) are rewritten as following,

$$\underline{y_{ij}} = \sum_{k=1,u_{kj}>0}^{p} \underline{X_{ik}} \cdot u_{kj} + \sum_{k=1,u_{kj}<0}^{p} \overline{X_{ik}} \cdot u_{kj}, \tag{3}$$

$$\overline{y_{ij}} = \sum_{k=1,u_{kj}>0}^{p} \overline{X_{ik}} \cdot u_{kj} + \sum_{k=1,u_{kj}<0}^{p} \underline{X_{ik}} \cdot u_{kj}. \tag{4}$$

Formally the following theorem is established:

Theorem 1. *Let X be an interval matrix defined by:*

$$X = \begin{pmatrix} [\underline{X_{11}}, \overline{X_{11}}] & \dots & [\underline{X_{1p}}, \overline{X_{1p}}] \\ \vdots & \ddots & \vdots \\ [\underline{X_{n1}}, \overline{X_{n1}}] & \dots & [\underline{X_{np}}, \overline{X_{np}}] \end{pmatrix}.$$

Let u_j be the j-th column vector of the matrix $U_{p\times p}$ and X_i be the i-th row of the matrix X. The (3) and (4) [Rodríguez 2000, Theorem 4.1.1] to calculate the maximum and minimum projections of the vectors $x_i \in X_i$ over u_j can be done by using the interval arithmetic to compute $X_i \cdot u_j$, that is:

$$\underline{y_{ij}} = \sum_{k=1,u_{kj}>0}^{p} \underline{X_{ik}} \cdot u_{kj} + \sum_{k=1,u_{kj}<0}^{p} \overline{X_{ik}} \cdot u_{kj}, \tag{5}$$

$$\overline{y_{ij}} = \sum_{k=1,u_{kj}>0}^{p} \overline{X_{ik}} \cdot u_{kj} + \sum_{k=1,u_{kj}<0}^{p} \underline{X_{ik}} \cdot u_{kj}. \tag{6}$$

Proof. Interval arithmetic establishes that given $a = [\underline{a}, \overline{a}]$ and $b = [\underline{b}, \overline{b}]$ intervals then:

$$a + b = [\underline{a} + \underline{b}, \overline{a} + \overline{b}], a - b = [\underline{a} - \overline{b}, \overline{a} - \underline{b}]$$

$$\forall c \in I\!R, ca = \left\{ [c\underline{a}, c\overline{a}] \text{ if } c \geq 0 \text{ and } [c\overline{a}, c\underline{a}] \text{ if } c < 0. \right.$$

Then $y_{ij} = X_i \cdot u_j = \sum_{k=1}^{p} X_{ik} \cdot u_{kj} = [\sum_{k=1}^{p} \underline{X_{ik} \cdot u_{kj}}, \sum_{k=1}^{p} \overline{X_{ik} \cdot u_{kj}}]$, where

$$X_{ik} \cdot u_{kj} = \begin{cases} \left[u_{kj} \cdot \underline{X_{ik}}, u_{kj} \cdot \overline{X_{ik}}\right] & \text{if } u_{kj} \geq 0, \\ \left[u_{kj} \cdot \overline{X_{ik}}, u_{kj} \cdot \underline{X_{ik}}\right] & \text{if } u_{kj} < 0. \end{cases}$$

Therefore,

$$\underline{X_{ik} \cdot u_{kj}} = \begin{cases} u_{kj} \cdot \underline{X_{ik}} & \text{if } u_{kj} \geq 0, \\ u_{kj} \cdot \overline{X_{ik}} & \text{if } u_{kj} < 0, \end{cases}$$

$$\overline{X_{ik} \cdot u_{kj}} = \begin{cases} u_{kj} \cdot \overline{X_{ik}} & \text{if } u_{kj} \geq 0, \\ u_{kj} \cdot \underline{X_{ik}} & \text{if } u_{kj} < 0. \end{cases}$$

This gives the following results:

$$\underline{y_{ij}} = \sum_{k=1, u_{kj} > 0}^{p} \underline{X_{ik}} \cdot u_{kj} + \sum_{k=1, u_{kj} < 0}^{p} \overline{X_{ik}} \cdot u_{kj},$$

$$\overline{y_{ij}} = \sum_{k=1, u_{kj} > 0}^{p} \overline{X_{ik}} \cdot u_{kj} + \sum_{k=1, u_{kj} < 0}^{p} \underline{X_{ik}} \cdot u_{kj}.$$

which correspond to the equations given in (5) and (6).

This result offered the basis for formulating a theorem for projection of a fuzzy number variable using fuzzy number arithmetic. Before presenting the theorem the following definitions are introduced:

- A trapezoidal fuzzy number Y is represented by $Y = (Y^{(1)}, Y^{(2)}, Y^{(3)}, Y^{(4)})$ and its membership function is defined by:

$$\mu_Y(x) = \begin{cases} 0 & \text{if } x < Y^{(1)}, \\ \frac{x - Y^{(1)}}{Y^{(2)} - Y^{(1)}} & \text{if } Y^{(1)} \leq x < Y^{(2)}, \\ 1 & \text{if } Y^{(2)} \leq x \leq Y^{(3)}, \\ \frac{Y^{(4)} - x}{Y^{(4)} - Y^{(3)}} & \text{if } Y^{(3)} < x \leq Y^{(4)}, \\ 0 & \text{if } Y^{(4)} < x. \end{cases} \tag{7}$$

- The support and core of the trapezoidal fuzzy number Y are defined as $\text{supp}(Y) = [Y^{(1)}, Y^{(4)}]$ and $\text{core}(Y) = [Y^{(2)}, Y^{(3)}]$.

- The α level for the trapezoidal fuzzy number Y, noted by Y_α, corresponds to $Y_\alpha = [Y^{(1)} + \alpha(Y^{(2)} - Y^{(1)}), Y^{(4)} - \alpha(Y^{(4)} - Y^{(3)})]$.
- The mean interval for a trapezoidal fuzzy number Y is defined by $\bar{E}(Y) = [E_*(Y), E^*(Y)]$ (Dubois 2006) where

$$E_*(Y) = \int_0^1 (\inf Y_\alpha)\, d\alpha = \int_0^1 (Y^{(1)} + \alpha(Y^{(2)} - Y^{(1)}) d\alpha = \frac{Y^{(1)} + Y^{(2)}}{2}$$

$$E^*(Y) = \int_0^1 (\sup Y_\alpha)\, d\alpha = \int_0^1 (Y^{(4)} - \alpha(Y^{(4)} - Y^{(3)}) d\alpha = \frac{Y^{(3)} + Y^{(4)}}{2}$$

- The middle point of mean interval for a trapezoidal fuzzy number Y is defined by $\frac{Y^{(1)}+Y^{(2)}+Y^{(3)}+Y^{(4)}}{4}$.

Theorem 2. *Let X be a trapezoidal fuzzy number matrix defined by:*

$$X = \begin{pmatrix} (X_{11}^{(1)}, X_{11}^{(2)}, X_{11}^{(3)}, X_{11}^{(4)}) & \dots & (X_{1p}^{(1)}, X_{1p}^{(2)}, X_{1p}^{(3)}, X_{1p}^{(4)}) \\ \vdots & \ddots & \vdots \\ (X_{n1}^{(1)}, X_{n1}^{(2)}, X_{n1}^{(3)}, X_{n1}^{(4)}) & \dots & (X_{np}^{(1)}, X_{np}^{(2)}, X_{np}^{(3)}, X_{np}^{(4)}) \end{pmatrix}.$$

Let u_j be a column vector in IR^p, X_i the i-th row of matrix X, $y_{i\,supp} = supp(X_i) \cdot u_j$, $y_{i\,core} = core(X_i) \cdot u_j$, $y_{i_\alpha} = (X_i)_\alpha \cdot u_j$ and y_i the trapezoidal fuzzy number defined as $(y_i^{(1)}, y_i^{(2)}, y_i^{(3)}, y_i^{(4)}) = (\underline{y_{i\,supp}}, \underline{y_{i\,core}}, \overline{y_{i\,core}}, \overline{y_{i\,supp}})$. Then

$$y_{i\,\alpha} = [y_i^{(1)} + \alpha(y_i^{(2)} - y_i^{(1)}), y_i^{(4)} - \alpha(y_i^{(4)} - y_i^{(3)})] = (X_i)_\alpha \cdot u_j.$$

The two components of $y_{i\,\alpha}$ can be written in another way, respectively as:

$$\begin{aligned} y_i^{(1)} + \alpha(y_i^{(2)} - y_i^{(1)}) = {} & \textstyle\sum_{k=1,u_{kj}>0}^{p} (X_{ik}^{(1)} + \alpha(X_{ik}^{(2)} - X_{ik}^{(1)})) \cdot u_{kj} \\ & + \textstyle\sum_{k=1,u_{kj}<0}^{p} (X_{ik}^{(4)} - \alpha(X_{ik}^{(4)} - X_{ik}^{(3)})) \cdot u_{kj}, \end{aligned} \tag{8}$$

$$\begin{aligned} y_i^{(4)} - \alpha(y_i^{(4)} - y_i^{(3)}) = {} & \textstyle\sum_{k=1,u_{kj}>0}^{p} (X_{ik}^{(4)} - \alpha(X_{ik}^{(4)} - X_{ik}^{(3)})) \cdot u_{kj} \\ & + \textstyle\sum_{k=1,u_{kj}<0}^{p} (X_{ik}^{(1)} + \alpha(X_{ik}^{(2)} - X_{ik}^{(1)}) \cdot u_{kj}. \end{aligned} \tag{9}$$

The proof has been omitted for briefness [Pacheco 2007, Theorem 3.2.2].

To plot the principal component plane and to show the membership function of an ordered pair of trapezoidal fuzzy numbers, a $3d$ (XYZ) space was used. The value of membership functions is represented on z axis and it is defined by $z_i = \mu_{(x_i,y_i)} = \min(\mu_{x_i}, \mu_{y_i})$, where (x_i, y_i) is an ordered pair of trapezoidal fuzzy numbers.

It was also demonstrated that the intervals PCA is a particular case of the `PCA-TF` [Pacheco 2007, Theorem 3.3.2] and thus, the classic PCA as well [Pacheco 2007, Theorem 3.3.1].

Algorithm – Principal Components Analysis for Trapezoidal Fuzzy Numbers (`PCA-TF`)

This algorithm is an extension to trapezoidal fuzzy numbers of the algorithm suggested by Rodríguez (2000)

Inputs

n = Number of individuals (number of rows of the data matrix).
p = Number of variables (columns of the data matrix).
X = Data table (matrix of trapezoidal fuzzy numbers).

Outputs

C = Principal components (trapezoidal fuzzy numbers matrix).
R = Correlation between variables and principal components (interval matrix).
CAL = Quality of representation of individuals (real number matrix).
CTR = Individual contribution to the components (real number matrix).
INR = Individual contribution to the total inertia (real number matrix).

Steps

1. Defuzzify X calculating the middle point of the mean interval. X^E represents defuzzified X.
   With $i = 1, \ldots, n$ and $j = 1, \ldots, p$, calculate:

$$X^E = ((X^E{}_{ij})) = \bar{E}(X) = (\bar{E}(X_{ij})) = \left(\left(\sum_{l=1}^{4} \frac{X_{ij}{}^{(l)}}{4}\right)\right).$$

2. Calculate the mean and standard deviation for the columns of the matrix X^E.
   With $i = 1, \ldots, n$ and $j = 1, \ldots, p$, calculate:

$$\bar{X}_j^E = \sum_{i=1}^{n} \frac{X^E{}_{ij}}{n}, \sigma_j{}^E = \left(\sum_{i=1}^{n} \frac{(X^E{}_{ij} - \bar{X}_j^E)^2}{n}\right)^{1/2}.$$

3. Calculate the matrix $Z = (z_{ij})$, where with $i = 1, \ldots, n$ and $j = 1, \ldots, p$, calculate: $z_{ij} = \frac{1}{\sqrt{(n)}} \frac{X^E{}_{ij} - \bar{X}_j^E}{\sigma_j^E}$.
4. Calculate the matrix $X^c = (X^c{}_{ij})$, where with $i = 1, \ldots, n$ and $j = 1, \ldots, p$, calculate: $X^c{}_{ij} = \frac{1}{\sqrt{(n)}} \frac{X_{ij} - \bar{X}_j^E}{\sigma_j^E}$.

5. Calculate the matrix $V = Z^t Z$.
6. Calculate the first q eigenvectors $u_1, u_2, \ldots, u_q$ of V and its associated eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q \geq 0$.
7. Calculate the principal components C_k. For $k = 1, \ldots, q$ do: $C_k = (X^c) \cdot u_k$, ${C^c}_k = Z \cdot u_k$
8. Calculate the eigenvector for the matrix ZZ^t using those of the $Z^t Z$ matrix. For $i = 1, \ldots, n$ y $j = 1, \ldots, q$ calculate: $w_{ij} = \frac{1}{\sqrt{\lambda_j}}(\sum_{k=1}^{n} z_{ik} \cdot u_{kj})$.
9. Calculate the correlations amongst variables and the principal components: For $k = 1, \ldots, q$ do: $PR_k = (X^c)^t w_k$, $R_k = (\text{supp}(PR_{ki} \cap [-1, 1])$,
10. Calculate the interpretation parameters:
    Individual representation quality for individual i in the factorial axis j

$$CAL(i, j) = CAL(X_i, u_j) = \frac{({C^c}_{ij})^2}{\sum_{i=1}^{p}(z_{ij})^2}.$$

    Contribution of individual i to the factorial axis inertia j

$$CTR(i, j) = CTR(X_i, u_j) = \frac{({C^c}_{ij})^2}{n * \lambda_j}.$$

    Contribution of the i to the total inertia

$$INR(i) = INR(X_i) = \frac{1}{\sqrt{n}} \frac{\sum_{i=1}^{p}(z_{ij})^2}{\sum_{i=1}^{p} \lambda_p}.$$

11 End of the algorithm.

3 Tests of the Performance of the PCA-TF Method

The performance of the suggested method has been tested on the example of two different data sets (student data set Denœux and Masson 2004, fruit juices data set Giordani and Kiers 2006) and the results were compared to those obtained by a similar method by Denœux and Masson (2004).

Let us consider first the hypothetical data set shown in Table 1, taken from Denœux and Masson (2004).

Applying the `PCA-TF`, the principal component plane and the correlation circle shown in Figs. 1 and 2 respectively, are obtained. The plot in Fig. 1 is generated by drawing the ordered pairs of fuzzy numbers (C_{ij}, C_{ik}), each of which represent the individual i, where j and k are the selected axes for visualization (in Figs. 1 and 2, $j = 1$ and $k = 2$). In this way the support rectangle, the core rectangle and the borders, where the membership function for ordered pair of trapezoidal fuzzy number experiences a change, can be examined. For the positioning of the individuals a traditional interpretation can be applied.

Table 1 Student dataset (Denœux and Masson 2004)

	M1	M2	P1	P2
TOM	15	fairly good	unknown	[14,16]
DAVID	9	good	fairly good	10
BOB	6	[10,11]	[13,20]	bien
JANE	very bad	very good	19	[10,12])
JOE	(0,0,2,6)	fairly good	[10,14]	14
JACK	(1,1,1,1)	[4,6]	9	[6,9]

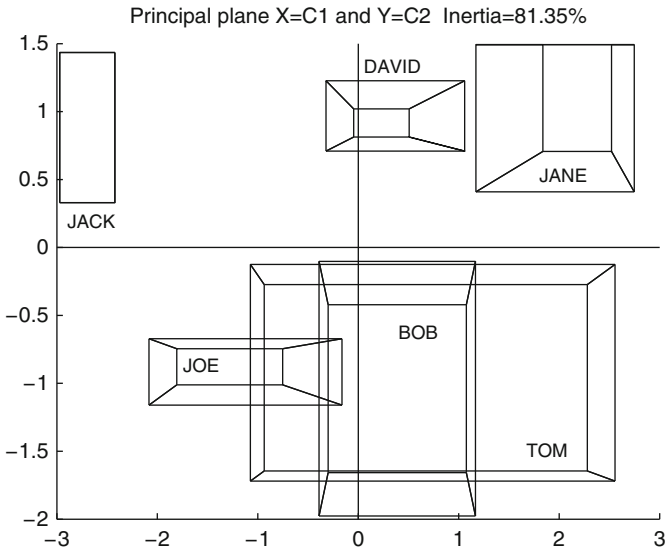


Fig. 1 Principal plane using the first two principal components

From the results shown in Figs. 1 and 2, it seems that the first principal component is related with the behavior in mathematics (variables M1 and M2), while the second principal component reflects the behavior in physics (variables P1 y P2). That is why Jane, who had the best grades in mathematics, appears in the upper right corner of principal component plane, while Jack having the worst grades – is at the left fringe. The second principal component has a high negative correlation with P2. Therefore, those with good grades on this variable are located in the inferior part of the principal component plane and viceversa. Besides it is noted that Jack is represented by a rectangle shape due to the fact that his data values are hard or "crisp" intervals, while the data for Jane consist of fuzzy numbers for the coordinates x and y in principal component plane, given that half of her grades are fuzzy numbers. Tom's grade in P1 is unknown and therefore it was represented by an interval for the whole range of grades variation, which yields the biggest support intervals for the coordinates among all the students.

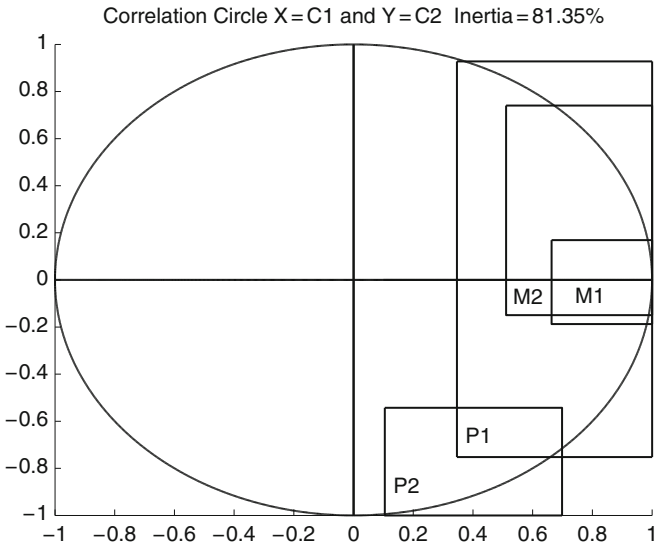


Fig. 2 Correlation circle using the first two principal components

In the correlations circle shown in Fig. 2, the correlation between $M1$ and $M2$ can be verified with first principal component, and that of $P2$ with the second principal component. $P1$ reflects the huge variation range in Tom's grade in this variable, as his data were modeled as [0,20] interval due to his unknown grade.

In the method proposed by Denœux and Masson (2004) (D&M-PCA) the standard iterative gradient descent is proposed, further on, there is no guarantee of the orthogonality of the axes, which is a clear weakness. Meanwhile classical PCA guarantees that the principal components are not correlated. Another drawback is that minimizing the function by means of the gradient descent algorithm a minimum local can be found instead of the absolute one, since it's a greedy algorithm. In `PCA-TF` the axes to which the matrix of fuzzy numbers are projected (which yields principal components) are orthogonal.

A comparison of the results of these two methods applied to the same data showed that they were similar, despite of the fact that in the method by Denœux & Masson the axes were not orthogonal.

`PCA-TF` was also applyed to fruit juices data set (Giordani and Kiers 2006) and the results were compared to `D&M-PCA` (method proposed by Giordani and Kiers 2006). The components have similar interpretations, but the axes to which the matrix of fuzzy numbers are projected are orthogonal, in the case of `PCA-TF` method (Pacheco 2007). The details has been omitted for briefness. Further work will be oriented to a simulation study to gain better knowledge on the performance of the method.

4 Conclusions

The study presented herein considered the theoretical aspects of several independent topics that all operate on interval variable type such as fuzzy numbers, interval arithmetic and symbolic data to propose a new method called PCA-TF. For that, an important relationship between the interval arithmetic and the formulas for projecting interval valued data was also proved (Theorem 1). Relationship between fuzzy numbers and intervals was already noted before by other authors.

The proposed PCA-TF method is an extension PCA to trapezoidal fuzzy numbers. The approach is developed in the context of symbolic data analysis for interval data type, applying the fundamental theory of operations on interval data and fuzzy number arithmetic. The performance of the method was tested on two data set of fuzzy numbers (student grades and fruit juices) and compared with D&M-PCA. PCA-TF method has an advantage of projecting over orthogonal axes and it yields the components which are trapezoidal fuzzy numbers. Both the classical PCA and the interval PCA are particular cases of PCA-TF. Further work will be oriented to a simulation study to gain better knowledge on the performance of the method.

Acknowledgements The authors dedicate this work to Suzanne Winsberg and thank the Costarican Institute of Electricity for the financial support of the participation in the IFCS 09 held in Dresden, Germany.

References

Bock, H.-H., & Diday, E. (Eds.). (2000). *Analysis of symbolic data: Exploratory methods for extracting statistical information from complex data*. New York: Springer.

Cazes, P., Chouakria, A., Diday, E., & Schektman, Y. (1997). Extension de l'analyse en composonantes principales à des données de type intervalle. *Review of Statistique Appliquée, XLV*(3), 5–24.

Denœux, T., & Masson, M. (2004). Principal component analysis of fuzzy data using autossociative neural Networks. *IEEE Transactions on Fuzzy Systems, 12*, 336–349.

Dubois, D. (2006). Possibility theory and statistical reasoning. *Computational Statistics & Data Analysis, 51*, 47–69.

Giordani, P., & Kiers, H. (2006). A comparison of three methods for principal component analysis of fuzzy interval data. *Computational Statistics & Data Analysis, 51*, 379–397.

Kaufman, A., & Gupta, M. (Eds.). (1991). *Introduction to fuzzy arithmetic: Theory and applications*. New York: Van Nostrand Rheinhold.

Lodwick, W., & Jamison, K. D. (2003). Special issue: Interfaces between fuzzy set theory and interval analysis. *Fuzzy Sets and Systems, 135*, 1–3.

Pacheco, A. (2007). *Análisis en Componentes Principales para Números Difusos Tipo Trapezoide*. San José: Universidad de Costa Rica.

Rodríguez, O. (2000). *Classification et Modéles Linéaires en Analyse des Données Symboliques*. Paris: Université Paris IX-Dauphine.

Factor Selection in Observational Studies – An Application of Nonlinear Factor Selection to Propensity Scores

Stephan Dlugosz

Abstract Observational studies have become a major research methodology in scientific disciplines where experiments are hard to perform. These most notably include the health sciences, social sciences and (macro-)economics. It is difficult to estimate treatment effects due to the non-randomised character of these studies. Propensity scores solve this problem to some extent by incorporating the control variables into one measure (Rosenbaum and Rubin 1983). An estimation of these propensity scores can be performed by any method of categorical regression (logit etc.). Nevertheless, estimated effects are very sensitive to the propensity score (Drake 1993; Heckman et al. 1998; Brookhart et al. 2006) and for this reason, the propensity score should be estimated with great care. In order to avoid a poor generalisation performance of the propensity score estimation, it is important to choose factors appropriately. Unfortunately, it is not possible to select factors during the estimation of the propensity score and independently from treatment calculation. In this paper, an integrated factor selection method is proposed, which considers the treatment effect estimation during the propensity score estimation.

1 Introduction

Empirical researchers in many scientific disciplines – such as sociology, economy, and epidemiology – are often faced with the problem that they cannot use randomised experiments if these are tied up with economic or ethical issues. Data for these kinds of research are usually obtained from observing subsets of the population. Unfortunately, often neither the sampling process nor the treatment assignment is randomised. In randomised experiments, simple comparisons between the treatment and the control group result in unbiased treatment effect estimates; but the same simple comparisons between the two groups usually result in an overestimation of the treatment effect in the presence of selection bias. This estimation

S. Dlugosz
ZEW Centre for European Economic Research, Mannheim
e-mail: dlugosz@zew.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_39,

bias is caused by confounding variables ($\mathbf{x}$) that are related to the treatment assignment (t) and to the outcome (y). Thus, a stratification of the sample according to all of these background variables is necessary. Then, we can estimate the treatment effect for each strata separately and calculate a weighted sum – usually according to the population frequencies of the strata – to approximate the average treatment effect.

This method works well as long as the number of variables – and therefore the number of strata – is small compared to the number of independent observations. This curse of dimensionality problem is usually solved by "compressing" the high dimensionality of the covariate space to a single dimension, which is the probability of being treated conditional on the covariates, i.e. the propensity score (PS) $P(t = 1|\mathbf{x})$. This dimensionality reduction is effective in the sense that stratification according to the PS yields the same correction for the biased estimation as exact matching (Rosenbaum and Rubin 1983). Unfortunately, the true PS is usually not known and has to be estimated; and – to make matters worse – the results are very sensitive to the specification of this PS estimation model (Drake 1993; Heckman et al. 1998). Flexible – non-parametric or at least non-linear – models should be used to take care of this. Unfortunately, these are very sensitive to high dimensional factor spaces and an appropriate method of factor selection is needed.

In their seminal paper, ROSENBAUM and RUBIN do not suggest any method of factor selection (Rosenbaum and Rubin 1983) and implicitly assume that every variable is relevant. Under this assumption, simple tests of independence (i.e. χ^2-tests) are suitable for selecting the relevant factors (cf. Rosenbaum 2002 for examples). Instead of testing independency, we should be more interested in testing high dependency as the example demonstrates. Often – in real-world studies – the PS is estimated via parametric methods like logistic regression; and factor selection for this regression model is done with standard parameter testing using bottom-up or top-down model construction (Hirano et al. 2003; D'Agostino 1998) and deriving stopping rules for the procedures (Judkins et al. 2007). Apart from the fact that these approaches are dependent on model choice, order of factor in- or exclusion, they are also algorithmically quite expensive for high numbers of factors. Another very appealing approach uses the following iterative procedure (Dehejia and Wahba 2002):

1. Start with a parsimonious model
2. Sort data according to estimated PS
3. Stratify observations based on PS
4. Test means of treated group and control group within stratum
   (a) If balanced: stop
   (b) If covariates are not balanced for some stratum, divide stratum
   (c) If a covariate is not balanced for many strata, add it or add interactions to the model and re-evaluate

This algorithm takes the special structure of the factor selection problem for PS models into account. Unfortunately, it is order-dependent – especially on the starting model – and thus it cannot ensure that only the most relevant factors are selected,

as it usually selects too many variables. Additionally, it is an expensive procedure because of the necessary re-evaluation of the whole PS model.

2 Theoretical Framework

With the causal relationships between the covariates, the treatment and the outcome (cf. Fig. 1) in mind, we can divide the covariates into four groups of variables according to their causal relationship to the treatment t and/or the outcome y with the following properties (cf. Brookhart et al. 2006; Dawid 1979) for the notation of conditional independence):

1. Variables are independent between the four groups, but not necessarily within the groups, i.e. $\mathbf{x}_0, \mathbf{x}_y, \mathbf{x}_{t,y}, \mathbf{x}_t$ are stochastically independent
2. Variables from group $\mathbf{x}_0$ are stochastically independent from the outcome and the treatment assignment, i.e. $y \perp\!\!\!\perp \mathbf{x}_0 \perp\!\!\!\perp t$
3. Variables from group $\mathbf{x}_y$ are independent from treatment, i.e. $\mathbf{x}_y \perp\!\!\!\perp t$

For notational simplification, let (y_0, y_1) denote the (possible) outcome variables for each unit in the control and in the treatment group with (y_1) and without (y_0) treatment. Obviously, y_1 is missing for the control group and y_0 for the treatment group (cf. Rosenbaum and Rubin 1983 for details).

4. Variables from group $\mathbf{x}_t$ are independent from the outcome, i.e. $\mathbf{x}_t \perp\!\!\!\perp (y_0, y_1)$
5. The model includes at least all relevant variables, i.e. strong ignorability of the treatment assignment is given (Rosenbaum and Rubin 1983): $(y_0, y_1) \perp\!\!\!\perp t|(\mathbf{x}_0, \mathbf{x}_y, \mathbf{x}_{t,y}, \mathbf{x}_t)$

This simple framework leads us directly to the following simple facts:

(a) $(y_0, y_1) \perp\!\!\!\perp t|(\mathbf{x}_y, \mathbf{x}_{t,y}, \mathbf{x}_t)$ (assumptions 1, 2 and 5)
(b) $(y_0, y_1) \perp\!\!\!\perp t|p(\mathbf{x}_{t,y}, \mathbf{x}_t)$ [PS (cf. Rosenbaum and Rubin 1983), assumption 3 and a]
(c) $\mathbf{x}_t \perp\!\!\!\perp (y_0, y_1)|(\mathbf{x}_{t,y}, \mathbf{x}_y)$ (assumptions 1, 3 and 4)

Obviously, the variables from group $\mathbf{x}_0$ are irrelevant for the PS model. Furthermore, the variables from $\mathbf{x}_y$ can naturally be excluded from the PS model, as they are independent from the treatment. Property (c) indicates that variables in $\mathbf{x}_t$ behave like an additional source of randomisation for the study. Thus, it should not be necessary to stratify by the variables from $\mathbf{x}_t$ and we can reduce the PS model by these variables. It seems that a restricted model $p(\mathbf{x}_{t,y}) = pr(t = 1|\mathbf{x}_{t,y})$ for the PS based on $\mathbf{x}_{t,y}$ is sufficient to control for the selection bias.

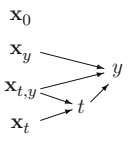


Fig. 1 Illustration of the causal framework

Proposition 1. *Under assumptions 1–5, it is true:*

$$(y_0, y_1) \perp\!\!\!\perp t | (\mathbf{x}_y, \mathbf{x}_{t,y})$$

Proof. Starting with $(y_0, y_1) \perp\!\!\!\perp t | (\mathbf{x}_y, \mathbf{x}_{t,y}, \mathbf{x}_t)$; by using c) and with Lemma 4.3 of Dawid (1979) it is true that $(y_0, y_1) \perp\!\!\!\perp (t, \mathbf{x}_t) | (\mathbf{x}_y, \mathbf{x}_{t,y})$ and with Lemma 4.2(i) of Dawid (1979) we get the result. □

Now we can show that $p(\mathbf{x}_{t,y})$ is sufficient to model the PS:

Corollary 1. *Under assumptions 1 to 5, it is true:*

$$\mathbf{E}\big[y_1 | p(\mathbf{x}_{t,y}), t = 1\big] - \mathbf{E}\big[y_0 | p(\mathbf{x}_{t,y}), t = 0\big] = \mathbf{E}\big[y_1 - y_0 | (\mathbf{x}_y, \mathbf{x}_{t,y}, \mathbf{x}_t)\big]$$

Proof. Application of Theorem 4 of Rosenbaum and Rubin (1983) by using Proposition 1 and afterwards making use of assumption 3. □

3 Factor Selection for Propensity Score Modelling

3.1 Non-linear Factor Selection

The idea of non-linear factor selection aims at the more general problem of factor selection for any kind of supervised (regression/classification) data analysis task for high dimensional categorical data. Two measures are needed for factor selection: A measure for the predictive power of a factor X on outcome Y to decide its relevance for the model in size and a measure of (dis-)similarity of the impact of two factors to deal with the association structure of the factors.

In general, a measure of predictive association could look like Dlugosz and Müller-Funk (2008)

$$A_D(Y|X) = 1 - L_X \left[\frac{D(Y|X)}{D(Y)} \right] \in [0, 1], \tag{1}$$

which is based on the relation of the dispersions of a-posteriori $D(Y|X)$ and a-priori $D(Y)$ distribution, which can be interpreted as the gain in precision of the prediction of outcome Y if the value for factor X is known. These relative gains have to be centred by a measure of location L_X over X (e.g. mean).

It is possible to deal with an ordinal Y appropriately with the help of an ordinal measure of dispersion (other variants are discussed in the same paper). Dlugosz and Müller-Funk (2008):

$$D_o(p) = L \left[\ldots, D\left(\sum_{i=1}^{r} p_i, \sum_{i=r+1}^{K} p_i \right), \ldots \right] \tag{2}$$

This measure is based on an ordinary measure of dispersion D for categorical variables Müller-Funk (2007).

A factor selection method which seeks for the factors with the largest predictive measure of association would neglect the associations between the factors, and most likely select a set of highly correlated factors. Therefore, a measure of difference in the impact of two factors V, W on the outcome has been proposed (Dlugosz and Müller-Funk 2008):

$$d(V,W) = \sum_{k,\ell} \left|\hat{\pi}_V(\cdot|k) - \hat{\pi}_W(\cdot|\ell)\right|_{TV} \hat{\pi}_{V,W}(k,\ell), \tag{3}$$

where $\hat{\pi}_V(\cdot|k)$ denotes the estimated probability of the outcome conditional on V having value k.

In order to render an ordinal measure of factor distances, the total variation in (3) can be replaced by a variant for ordinal variables

$$|\pi_1 - \pi_2|_{oTV} = \frac{1}{C-1} \sum_{c}^{C-1} \left| \sum_{i \le c} \pi_1(i) - \sum_{i \le c} \pi_2(i) \right| + \left| \sum_{i > c} \pi_1(i) - \sum_{i > c} \pi_2(i) \right|,$$

whose construction principle is similar to that of the ordinal dispersion measure (cf. Dlugosz and Müller-Funk 2008 for details).

The distances between the factors are used in a hierarchical cluster algorithm with an inhomogeneity minimising objective function – like 'complete linkage' – to identify groups of similar factors. In a subsequent data analysis, one representative factor from each of the groups should be selected according to the predictive association of the factor with the target variable.

3.2 Factor Selection for the Propensity Score Model

The following two conditions should hold for variables within the PS model, i.e. for variables, that belong to $\mathbf{x}_{t,y}$:

1. The relevant variables should be close to the treatment in a factor selection for a model that explains the outcome with the help of the independent variables and the treatment (treated as one of the covariates).
2. The relevant variables should be close to the outcome in a factor selection for a model that explains the treatment assignment with the help of the independent variables and the outcome (treated as one of the covariates).

Based on this idea, the following algorithm is proposed:

1. calculate the two dendrograms of $(X,T) \to Y$ and $(X,Y) \to T$
2. use tests to identify $\mathbf{x_0}$

3. identify variable groups:

- $\mathbf{x}_y \subset A \subset \mathbf{x}_y \cup \mathbf{x}_{t,y}$ through distance to treatment
- $\mathbf{x}_t \subset B \subset \mathbf{x}_{t,y} \cup \mathbf{x}_t$ through outcome dendrogram (covered by treatment)
- $\mathbf{x}_y \subset C \subset \mathbf{x}_y \cup \mathbf{x}_{t,y}$ through treatment dendrogram (covered by outcome)
- $\mathbf{x}_t \subset D \subset \mathbf{x}_{t,y} \cup \mathbf{x}_t$ through distance to outcome

$\rightarrow$ identification through intersections, i.e.
$\mathbf{x}_y \subset A \cap C$ and $\mathbf{x}_t \subset B \cap D$ and then $\mathbf{x}_{t,y} \supset (A \cup C) \cap (B \cup D)$

4 Example

This example is based on an artificial data set describing an evaluation of a labour market programme from 1976 in the US (LaLonde 1986). The treatment is programme participation and the measured outcome is real income for the year 1978. Possible covariates are age, education (in years), genetic origin, marital status, high-school degree, real income for the years 1974 and 1975 and longer unemployment phases in 1974 and 1975.

The factor selection method requires categorical data. For this reason, some of the variables have to be categorised. For better results, the categorisation procedure follows two principles: (1) use only a few categories to speed up calculation time and sufficient number of cases per cell; (2) choose boundaries by maximising variance on categorised variables to maximise the power of the method. Additionally, the variables 'hispanic' and 'black', which are mutually exclusive, are combined to one variable called 'color'.

In a first step, the measures of association A_D and A_{D_o} as well as tests on factor relevance, i.e. simple Wald and likelihood-ratio tests for 'linear' dependence (cf. Dlugosz and Müller-Funk 2008) and χ^2-tests for independence (cf. Rosenbaum and Rubin 1983) have been computed (cf. Table 1 for corresponding p-values).

Table 1 LaLonde data: Predictive associations and p-values of some factor relevance tests

	Outcome				Treatment		
Factor	A_D	A_{D_o}	LR	χ^2	A_D	Wald	χ^2
Age	0.0311854	0.0308896	0.442	0.671	0.0209229	0.785	0.904
Education	0.0179272	0.0247675	0.131	0.095	0.0367440	0.400	0.010
Color	0.0153731	0.0306167	0.011	0.001	0.0096511	0.129	0.115
Married	0.0021067	0.0011027	0.938	0.486	0.0016722	0.467	0.369
No degree	0.0046661	0.0062814	n.a.	0.031	0.0233209	0.017	0.002
Real income 74	0.0077462	0.0056868	0.365	0.553	0.0037928	0.374	0.810
Real income 75	0.0087970	0.0107277	0.976	0.832	0.0206913	0.655	0.107
Real income 78	1	1	–	–	0.0229913	0.041	0.062
Unemployed 74	0.0019883	0.0023366	n.a.	0.598	0.0023383	0.736	0.330
Unemployed 75	0.0024489	0.0029584	n.a.	0.418	0.0080206	0.546	0.070
Treatment	0.0058111	0.0086944	0.053	0.065	1	–	–

The results from Table 1 imply – χ^2-tests for both aspects are significant at the 10% level – that the simple factor selection methodology inspired by Rosenbaum and Rubin (1983) would yield the variables 'education', 'color' and 'no degree'. In particular the exclusion of 'age' is not plausible from an economic point of view. Furthermore, due to its exemplary character, we could not expect the LALONDE data to contain irrelevant variables in advance. This is why the variable 'random' – describing a binary random variable – has been added to the data set. Following the idea of 'grouping' the variables according to their 'unique' information on the target variable, variables that do not have a high impact on the outcome should be 'close' to 'random'. Please note that this is neither a test nor sufficient, it is simply an illustration of the methodology.

Applying the techniques proposed in this paper yields the dendrograms Fig. 2. The first dendrogram shows the impact of the factors on the outcome and the second dendrogram shows their impact on the treatment.

We define the four classes of factors by dividing the dendrograms at 'treatment' (groups A and B) and 'real income 78' (groups C and D), respectively: A includes 'education', 'color', and 'age'; B is 'income 75', 'no degree', 'income 74', 'unemployed 74', 'unemployed 75', and 'married'; C has 'income 75', 'income

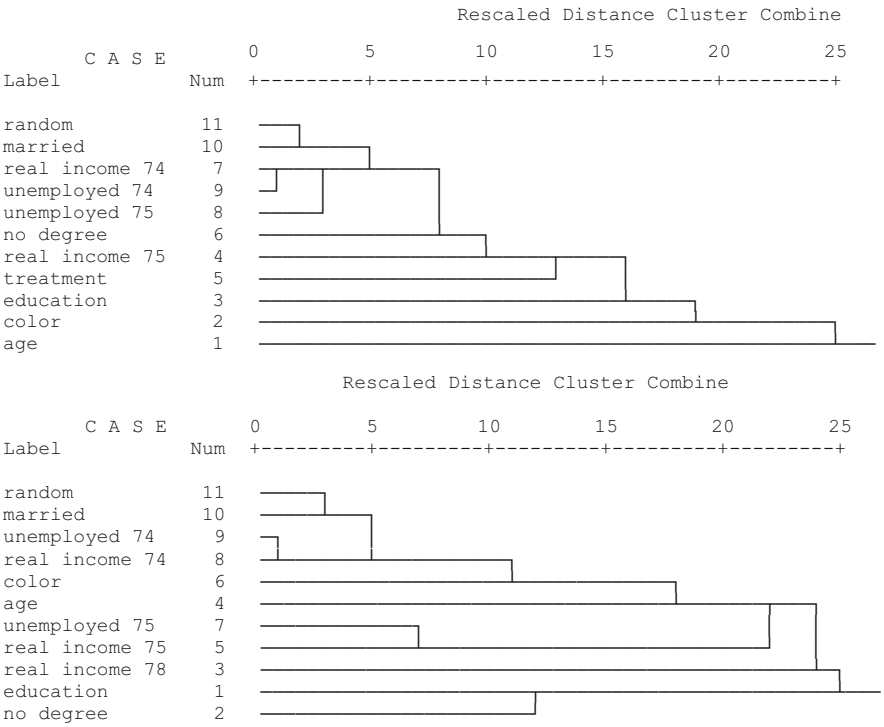


Fig. 2 Dendrograms of clustered covariates

74', 'unemployed 74', 'unemployed 75', 'color', 'age', and 'married'; and D is 'no degree', and 'education'. We identify 'married' as a less important variable (group x_0) by looking at the 'random' variable. The other three groups are identified by the intersections and result in (a) $\hat{\mathbf{x}}_y = A \cap C$ with 'color', 'age'; (b) $\hat{\mathbf{x}}_t = B \cap D$ with 'no degree' and (c) $\hat{\mathbf{x}}_{t,y}$ with 'income 74', 'income 75', 'unemployed 74', 'unemployed 75', and 'education'.

This result demonstrates the power of the new methodology as it leads to well-interpretable conclusions. The covariates 'age' and 'color' are known to be important in economic models for income (Cahuc and Zylberberg 2004). It is plausible that they do not matter in explaining the assignment to the labour market programme as well as the participation decision to attend the programme of the chosen participants. Also, the variable 'no degree' is more relevant for programme assignment than for labour market outcomes, as it is just a ticket for participation and does not determine your wage (it does not, in fact, determine more than 'education'). Marriage does not seem to be important at all, which is also plausible. The other variables reflect the employment history and the educational background. These are relevant for the labour market outcome and also for the assignment to labour market programmes as only the unemployed and poorly educated participate in such programmes.

The new method is superior to older methods in terms of computation time, it deals with the special structure of the factor selection problem for PS models, and it is non-parametric. Although the variables have to be categorised to be handled, this explorative technique produces plausible results (in a popular example) and the information loss through categorisation can be reduced by using measures for ordinal data.

Acknowledgements The author is supported by the German Research Foundation through the "Statistical Modelling of Errors in Administrative Labour Market Data" grant.

References

Brookhart, A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*(12), 1149–1156.

Cahuc, P., & Zylberberg, A. (2004). *Labor economics*. Cambridge, MA: MIT Press.

D'Agostino, R. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17*, 2265–2281.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society B, 41*(1), 1–31.

Dehejia, R. H., & Wahba, S. (2002). Propensity score matching methods for non-experimental causal studies. *Review of Econometrics and Statistics, 84*(1), 151–161.

Dlugosz, S., & Müller-Funk, U. (2008). Predictive classification and regression trees. *Proceedings of the 32nd Annual Conference of the Gesellschaft für Klassifikation e.V.*

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics, 49*, 1231–1236.

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica, 66*, 1017–1098.

Hirano, K., Imbens, G., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica, 71*, 1161–1189.

Judkins, D. R., Morganstein, D., Zador, P., Piesse, A., Barrett, B., & Mukhopadhyay, P. (2007). Variable selection and ranking in propensity scoring. *Statistics in Medicine, 26*, 1022–1033.

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review, 76*, 604–620.

Müller-Funk, U. (2007). Measures of dispersion and cluster-trees for categorical data. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme and R. Decker (Eds.), *Data analysis, machine learning and applications* (pp. 163–170). Berlin: Springer.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York: Springer.

Rosenbaum, P. R., & Rubin D. (1983). The central role of the propensity score in observational studies for causal effekts. *Biometrika, 70*(1), 41–55.

Nonlinear Mapping Using a Hybrid of PARAMAP and Isomap Approaches

Ulas Akkucuk and J. Douglas Carroll

Abstract Dimensionality reduction aims to represent higher dimensional data by a lower-dimensional structure. A well-known approach by Carroll, Parametric Mapping (PARAMAP) (Shepard and Carroll 1966) relies on iterative minimization of a loss function (called kappa or "κ") measuring the smoothness or continuity of the mapping from the lower dimensional representation to the original data. The approach was resuscitated recently with important algorithmic modifications (Akkucuk 2004; Akkucuk and Carroll 2003, 2006). However improved, the approach still involved the need to make a large number of random starts. In this paper we discuss the use of a variant of the Isomap method (Tenenbaum et al. 2000) to obtain a starting framework, consisting of a core set of landmark points. These core set of landmark points are used to construct a rational start for running PARAMAP algorithm only once. Since Isomap is faster and less prone to local optimum problems than PARAMAP, and the iterative process involved in adding new points to the configuration will be less time consuming (since only one starting configuration is used), we believe the resulting method should be better suited to deal with large sets of realistically based data, and more inclined to obtain a satisfactory solution in reasonable time.

1 Introduction

Dimension reduction techniques involve algorithms aiming to find a lower dimensional structure which preserves the characteristics of the higher dimensional input configuration. Nonlinear mapping techniques specialize in structures where the relationship between the coordinates is highly nonlinear. Prime examples to nonlinear surfaces are spheres, tori and spirals such as those shown in Fig. 1. PARAMAP (Shepard and Carroll 1966) approach and Isomap (Tenenbaum et al. 2000) approach are two methods that deal with such nonlinear surfaces. A comparison of PARAMAP

U. Akkucuk (✉)
Department of Management, Bogazici University, Istanbul, Turkey
e-mail: ulas.akkucuk@boun.edu.tr

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_40,

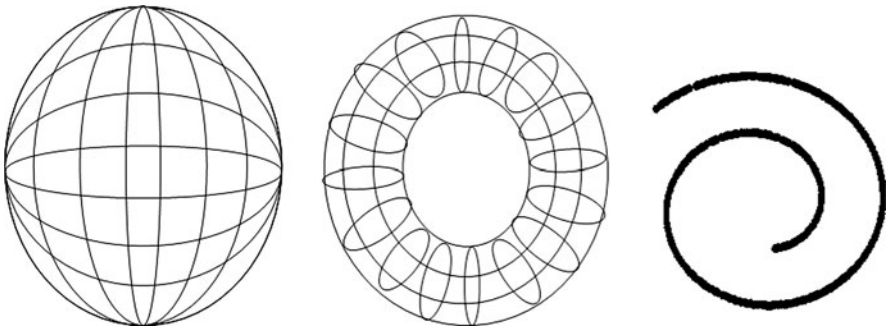


Fig. 1 Representative nonlinear manifolds (*from left to right*): A sphere in three dimensions, a torus in three dimensions and a spiral in two dimensions

and Isomap algorithms was provided by Akkucuk and Carroll (2006). In this former paper, solutions provided by Isomap and PARAMAP were compared using various performance criteria. The main conclusion of Akkucuk and Carroll was that Isomap was not designed to handle closed surfaces and also open surfaces with insufficient point density. PARAMAP could handle closed surfaces well but computation time was prohibitive.

In this paper our purpose is to propose a hybrid technique that will combine important features of Isomap and PARAMAP. The new algorithm will use Isomap as a precursor to PARAMAP, hence the resulting algorithm will not require different random starts, but will require a number of quick Isomap runs followed by one PARAMAP run. The initial Isomap run will result in either an incomplete or complete solution (in the sense of the lower dimensional embedding containing the same number of points as the higher dimensional input configuration). If the solution is incomplete, the points in the lower dimensional solution provided by Isomap will be regarded as the landmark points. Our proposed algorithm will create the starting configuration from the incomplete solution containing the landmark points and use this "rational" starting configuration as input to PARAMAP. This paper will proceed as follows: Sect. 2 will review the salient technical features of PARAMAP and Isomap. Section 3 will present the new proposed hybrid approach combining Isomap and PARAMAP. Section 4 will present the results of running the algorithm on different experimental configurations, involving nonlinear manifolds. Finally Sect. 5 will give the conclusions and directions for future research.

2 PARAMAP and Isomap Algorithms

The technical features of the two algorithms were discussed extensively in a number of previous papers, especially in Akkucuk and Carroll (2006). Here we will provide a summary of the essential terminology for understanding the arguments we make in this paper.

2.1 The PARAMAP Algorithm

The PARAMAP algorithm attempts to find a lower dimensional configuration taking as input a higher dimensional configuration. In doing so PARAMAP tries to minimize a measure of continuity called "*kappa*" or abbreviated as κ, that relates the lower dimensional configuration to the higher dimensional configuration. The *kappa* function is computed by using the input dissimilarities (to be called δ_{ij}) computed from the higher dimensional input configuration and the distances computed from a lower dimensional configuration (to be called d_{ij}). The *kappa* function is computed as shown in (1):

$$\kappa = \sum_{i \neq j} \frac{\delta_{ij}^2}{d_{ij}^4} \Bigg/ \left[\sum_{i \neq j} \frac{1}{d_{ij}^2} \right]^2 \tag{1}$$

This function measures the degree of continuity inversely; hence smaller values indicate that the higher dimensional structure is well represented by the lower dimensional solution. This function is minimized using a gradient based approach with various parameters to be specified by the user. The most important parameters being the number of random starts to use, the maximum number of iterations permitted and the length of gradient stopping criterion.

It is worthwhile to make a note here about the values of κ that are reported in this paper. In order to standardize κ it is necessary to consider the condition in which the δ_{ij} and d_{ij} are perfectly correlated. If κ is multiplied by a constant that is only a function of the δ_{ij}, in particular if this constant is chosen such that it is the inverse of κ when $\delta_{ij} = d_{ij}$, then κ will be standardized such that the lowest possible value is 1. This naturally has no effect on the minimization process, but results in an easier interpretation of κ.

2.2 The Isomap Algorithm

The Isomap algorithm (Tenenbaum et al. 2000) uses a shortest path method to create the matrix of the "geodesic" distances between the points lying on a nonlinear manifold, rather than the straight line Euclidean distances. Before running the shortest path algorithm, the nearby points are connected by using either of the two options, namely by connecting the "k nearest neighbors" or "neighbors closer than a small value ε". After this connection step, the shortest path procedure computes the shortest path distances between all the points and then a classical metric MDS step creates the lower dimensional solution. After the lower dimensional soution is found, the Isomap program computes a measure of mapping performance called the "residual variance" which we abbreviate by *RV*. This measure is equal to 1 – (the squared correlation between the geodesic distances and the distances computed from the lower dimensional embedding). Naturally, the closer the value is to "0" the better the

Isomap mapping performance since the true geodesics are reflected in the classical MDS solution.

The solutions and the *RV* values change depending on the choice of the parameter k or ε. It is also clear that connected (i.e., the full set of points in the input configuration are represented in the solution) or disconnected (i.e., a subset of the full set of points in the input configuration is represented in the output configuration) solutions may occur depending on the selection of the value of the parameters of Isomap. When the k nearest neighbors option is used it is impossible to end up with disconnected solutions but when the other option (neighborhood size epsilon) is used it is possible that below a threshold value some points will not be able to connect and will be left out of the solution.

The proper selection of the neighborhood size parameter (ε) has been discussed in the literature and one suggestion was increasing the neighborhood size at regular intervals, starting from a low value, and observing *RV*. It was reported that the *RV* value starts decreasing slowly than at a particular point experiences a sudden jump upwards (Balasubramanian et al. 2002). It may thus be wise to stop and choose the solution with the minimum *RV* right before the sudden jump. In our experiments we also observed a similar phenomenon. Also, we observed that in some of those cases the graph was not fully connected and the total number of points in the solution did not equal the number of points in the input configuration. The main reason in the sudden deterioration is that as ε increases some points on opposite sides of the manifold are connected (an act known as "short-circuiting" or "cross-branching", similar to for example connecting two opposite points on a sphere by going inside the sphere rather than going over the surface of the sphere) and this perturbs the actual geodesic distances. The solution strategy we offer in this paper is taking this partially complete solution (but otherwise geometrically sound) solution and use it as an input to PARAMAP program hence eliminating the need to do many random starts with the computationally expensive algorithm.

2.3 Evaluation of the Mapping Results

The values of κ and *RV* could be used to judge the quality of the mapping that takes place by either of the algorithms or the hybrid algorithm that we are proposing. Another possible measure is called the "Rate of Agreement in Local Structure" also abbreviated as A or "agreement rate" (Akkucuk and Carroll 2006). The agreement rate takes a certain neighborhood size (this is not related to the neighborhood size used in Isomap and in our applications we have taken this number as 5) and computes the percentage of points that are in the neighborhood of the same points both in the input configuration and in the output configuration (order is not important). The agreement rate has a maximum theoretical value of 100% and a minimum of 0%. In this paper κ, *RV* and A will be used to evaluate the mapping performance of various results. The ideal (or best) values are respectively 1.00, 0.00 and 100.00 (the last one being expressed generally as a percentage).

3 PARAMAP-Isomap Hybrid Approach

The main difference of the Isomap approach over PARAMAP is its ability to converge in reasonable time. The hybrid algorithm takes advantage of this feature of Isomap and uses a configuration obtained by Isomap as an initial starting configuration which will be further fine tuned by PARAMAP. The PARAMAP algorithm will only be run once such that the additional costs of running many random starts will be eliminated. In certain Isomap solutions the number of data points may be less than the number of data points in the input configuration (an incomplete start). In certain solutions the number of data points in the solution will be equal to that of the input configuration (a complete start). If the start is incomplete, the so called "core" set of landmark points will define the "interior" of the ultimate total configuration we aim to fit, and then new points will be added based on their proximity to landmark points in this initial framework. As new points are added to this core set of landmark points, the complete framework is gradually built via an evolutionary process by which points farther and farther out in the periphery of the ultimately defined structure are incorporated into what will be this final structure. The iterative process underlying PARAMAP also will tend to change the interrelationships among points in the initial framework, so that the final configuration may bear relatively little relationship to this initial structure. The proposed algorithm will be composed of the initial "Isomap Preprocessing Step" and the second and final step "Mapping the Holdout Points".

3.1 Isomap Preprocessing Step

In this step of the hybrid procedure, the input configuration will be subject to analysis via the Isomap algorithm. There will be more than one run since it is essential here to try different values of the parameter ε and compare the residual variance (RV) values reported by the Isomap software. Since Isomap converges vey fast the few additional runs made at this step need not increase the total computational burden.

As already mentioned, Isomap takes as input a user specified parameter which can be either one of the options "k" or "ε". The first option connects to each point the k nearest neighbors hence constructs the neighborhood graph. The latter option connects the points that are less than distance of ε to each point and hence constructs the neighborhood graph. It is evident that with the first option disconnected (incomplete) solutions are not possible while with the second option we might have some disconnected solutions and this actually forms the basis of our algorithm.

When selecting different values of ε one sensible technique is to start from "0" and increment in fixed step sizes. The fixed step size should be small enough to permit disconnected solutions. In some data sets the solutions will not be disconnected due to sufficient point density. The step size can also be formulated as a function of the interpoint distances. In this work we used a step size as small as

0.05. This was sensible with the data sets we used but may not work with other data sets. When the *RV* values for different step sizes are recorded, at some point the RV values should degenerate (increase rapidly). This will be due to an effect known as "cross branching", i.e., the step size is too much and the nearest neighbor turns out to be a point on the other side of the manifold. The starting configuration for the PARAMAP algorithm will be selected to be the one just before the degeneration occurs.

3.2 Mapping the Holdout Points

In this stage of the algorithm the holdout points are mapped in one step. The entire set of holdout points are first located at the points they are closest to in the input configuration. Then one PARAMAP run is taken to find the solution. During the PARAMAP run the positions of the landmark points are free to vary as are the positions of the holdout points.

4 Results on the Experimental Configurations

There are three experimental configurations on which we test the hybrid algorithm. The sphere with 62 regularly spaced points (composed of 5 parallels, 12 meridians and 2 poles), the Swiss Roll with 1,000 points (this data set is essentially a spiral generalized to three dimensions and was used in some of the data analysis in Tenenbaum et al. 2000), and a sphere with 1,000 points randomly generated to fall on the surface but with unequal spacing between them.

4.1 Sphere with 62 Regularly Spaced Points

This is a very special case where the Isomap algorithm results in a complete projection if the parameter ε is selected such that the 62 points are connected. The reason for this is that the Isomap procedure results in the points opposite to each other on the surface of the sphere to be mapped on to each other. The agreement rate for this solution is as low as 48.71%. If the parameter ε is selected such that the *RV* measure is minimized then this results in 12 points being mapped but with superior mapping performance (100% agreement rate). Although this particular case does not have a large number of points it may be an indicator that the algorithm is promising. According to a previous result [3], the best PARAMAP solution has an agreement rate of 79.35% and κ value of 1.13. This particular configuration in two dimensions is shown in Fig. 2. This solution is what we believe to be the optimal solution and has been obtained by doing 300 random starts.

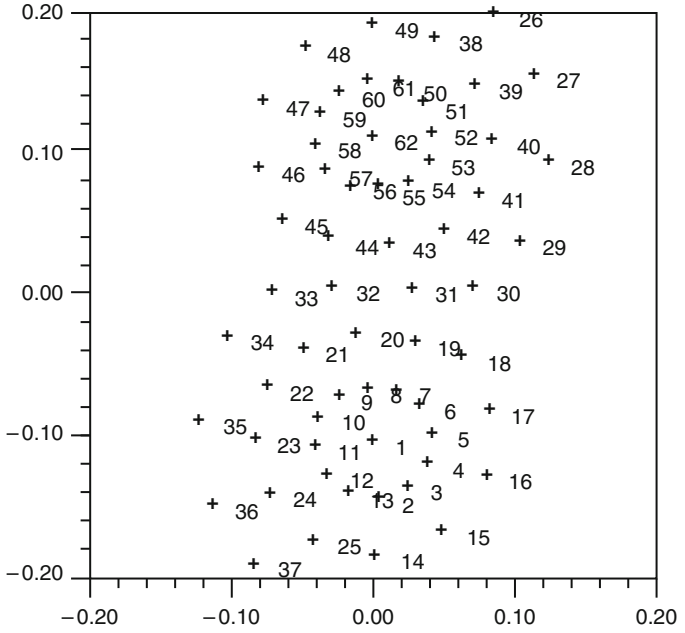


Fig. 2 PARAMAP globally optimal solution to the 62 points regularly spaced on the sphere

Table 1 Isomap preprocessing step results on the regular sphere with 62 points

ε	Number of connected components	Size of largest component	*RV*
0.30	40	12	0.050
0.35	40	12	0.050
0.40	18	12	0.050
0.45	18	12	0.050
0.50	***18***	***12***	***0.050***
0.55	1	62	0.320

When the hybrid algorithm is used, the Isomap common step procedure results in 12 connected points. Indeed this turns out to be the points 50–61, which we may refer to as the northernmost parallel. Table 1 shows the changes in *RV* values when the Isomap algorithm is run varying ε between 0.30 and 0.55. The point, right before the sudden jump in *RV* value, is shown in bold and italics. In order to be able to compare the solutions with the solutions obtained using PARAMAP alone we will provide here the agreement rate and kappa values of the best solution by PARAMAP. When the remaining $62 - 12 = 50$ points are mapped in at the same time, we obtain the solution given in Fig. 3. This solution has an agreement rate of 75.48% and a κ value of 1.21. The results indicate that the hybrid algorithm cannot reach the global optimum but produces satisfactory solutions in terms of preserving local structure. When Isomap is run alone, the users need to contend with a projection or

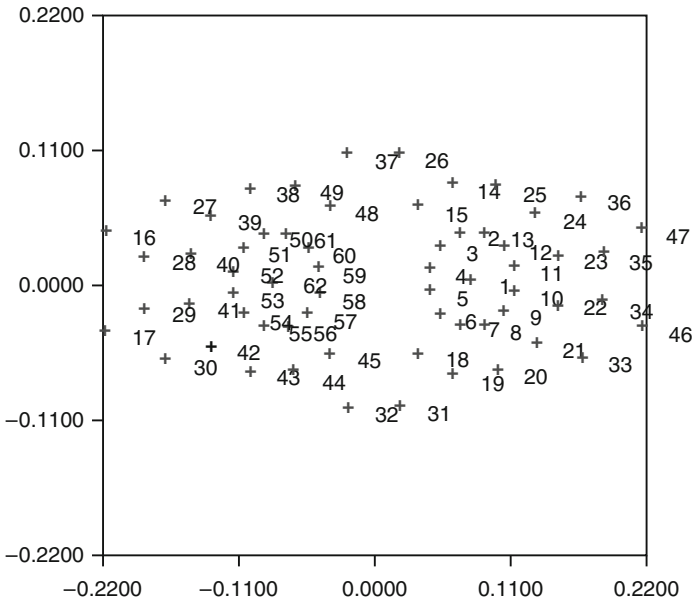


Fig. 3 Mapping of the 62 points on the sphere obtained by the proposed hybrid algorithm

a solution where a large fraction of the points are missing. In terms of run times, the hybrid algorithm is obviously superior to using only PARAMAP since the hybrid algorithm requires one run, and no random starts. However the flexibility of doing many random starts is foregone.

4.2 Sphere with 1,000 Points

In this configuration, the points fall exactly on the surface of a sphere, i.e., the Euclidean distance between all points and the center of the sphere is 1. Yet the spacing between the points is not regular. The preprocessing step this time results in a skeleton solution with 431 points as shown in Table 2 (in bold and italics). The skeleton configuration with 431 points is used as input to a single run of PARAMAP. Using the hybrid algorithm results in $\kappa = 1.146$ and $A = 73.01\%$. These values indicate once again that a good mapping has taken place. In Akkucuk (2004), using another method of selecting landmark points the same data set was mapped using PARAMAP. The former landmark selection procedure chose a fixed number of landmark points and did multiple random starts with the smaller set. After choosing the best solution with respect to κ, and placing the holdout points on top of the points they are closest to, one PARAMAP run was performed. This second phase of the former algorithm was essentially similar to the hybrid algorithm we are describing here. The two levels of landmark points experimented before resulted in different solutions. Thirty-two points resulted in a κ value of 1.295 and 100 landmark

Table 2 Isomap preprocessing step results on the regular sphere with 1,000 points

ε	Number of connected components	Size of largest component	RV
0.05	623	63	0.032
0.10	222	227	0.059
0.15	***31***	***431***	***0.034***
0.20	2	997	0.197
0.25	1	1000	0.211
0.30	1	1000	0.224
0.35	1	1000	0.229

points resulted in a κ value of 1.073. Also respectively the agreement rates were 75.84% and 77.88%. Our hybrid algorithm performs quite well compared to the other landmark selection procedure. It should also be noted that the previous landmark procedure requires many PARAMAP random starts for the landmark points (if not for the whole set of points) and one PARAMAP run for the complete set, our hybrid procedure requires only one PARAMAP run for the complete set. This results in savings in computational time.

4.3 Swiss Roll with 1,000 Points

This shape is the three-dimensional generalization of a spiral in two dimensions. In this data set the common step results in a fully connected solution. Further runs by PARAMAP do not result in any improvements. Agreement rate is 78.74% and does not improve further. This is an example of a case where the Isomap algorithm produces a solution that is sound both with respect to the "global" criterion that Isomap tends to minimize and the "local" criterion that PARAMAP tends to optimize.

5 Conclusion

We believe this hybrid approach is a very useful data analytic tool and can deal with either open surfaces with insufficient point density or with closed surfaces with a large number of points. The former case is difficult for Isomap since the resulting solution will not be complete. In this case the proposed hybrid algorithm can be used to map in the remaining points. In the latter case running different random starts with PARAMAP may be costly, so the hybrid algorithm can be used and only one PARAMAP run is necessary. In cases where Isomap produces a complete solution with a superior residual variance, then PARAMAP can be used to test the optimality of the solution. PARAMAP generally results in no further improvements in such a case since Isomap apparently produces the globally optimal solution also according to the PARAMAP criterion κ.

One possible avenue of further research is exploring the use of mapping the hold-out points in batches instead of mapping in all points simultaneously. In this case the optimal selection of the batch size is of concern. Another important issue for further research opportunity is the study of using the agreement rate as a stopping criterion in the initial Isomap procedure rather than the *RV* stopping criterion. One other practical improvement to the algorithm would involve integrating the two approaches into a single program (perhaps coded in Matlab or R), thus creating a single package with many options. Finally, an important problem with the hybrid algorithm is the inability to test different random starts. This problem can be partially solved by adding relatively small random perturbations to the final solution and running PARAMAP a number of times from such differently perturbed configurations. This may have the effect of enabling finding a nearby globally optimal solution.

References

Akkucuk, U. (2004). Nonlinear mapping: Approaches based on optimizing an index of continuity and applying classical metric MDS to revised distances (Doctoral dissertation, Rutgers University, Newark, NJ, 2004). University Microfilms No. AAT 3148774.

Akkucuk, U., & Carroll, J. D. (2003). *Nonlinear mapping: Approaches based on optimizing an index of continuity and applying classical metric mds to revised distances.* Paper presented at DIMACS working group meeting, Tallahassee, Florida.

Akkucuk, U., & Carroll, J. D. (2006). PARAMAP vs. isomap: A comparison of two nonlinear mapping algorithms. *Journal of Classification, 23*, 221–254.

Balasubramanian, M., Schwartz, E. L., Tenenbaum, J. B., DeSilva, V., & Langford, J. C. (2002). The isomap algorithm and topological stability. *Science, 295*, 7a.

Shepard, R. N., & Carroll, J. D. (1966). Parametric representation of nonlinear data structures. In P. R. Krishnaiah (Ed.), *Multivariate Analysis* (pp. 561–592). New York: Academic Press.

Tenenbaum, J. B., DeSilva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science, 290*, 2319–2323.

Dimensionality Reduction Techniques for Streaming Time Series: A New Symbolic Approach

Antonio Balzanella, Antonio Irpino, and Rosanna Verde

Abstract A growing number of applications generates massive streams of data which are on-line collected and potentially unbounded in size.

To cope with the high dimensionality of data, several strategies for dimensionality reduction have been proposed.

In this paper we introduce a new approach to represent an append only data stream into a reduced space. The main aim is to transform a real valued data stream into a string of symbols. The string includes a level component and a shape component allowing to get a better representation of data while maintaining a strong compression ratio.

1 Introduction

A growing number of applications in several disciplines including telecommunications, climate monitoring, security, wide-area sensor networks, generates massive streams of data over time.

In recent years there has been an increasing interest for data mining on such kind of data and many works have been proposed for clustering, classifying, detecting frequent patterns, approximating, data streams.

Streaming data are on-line collected and are potentially unbounded in size, this poses several computational and mining challenges:

- It is no longer possible to process data efficiently by using multiple passes;
- Data after processing are discarded or archived and become not easily available anymore;
- Memory resources are reduced withărespect to the amount of data to process;

A. Balzanella (✉)
University of Naples Federico II, Italy
e-mail: balzanella2@alice.it

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_41,

Since data stream analysis is performed using a bounded amount of memory it is not always possible to produce exact answers, so high-quality approximate ones are usually acceptable.

Moreover, often, in data stream mining tasks it is not possible to produce a new answer just when a new observation is seen, this is because the time required for computing answers can be bigger than the inter-arrival time among observations. In order to deal with this problem strategies performing dimensionality reduction, based on data sampling, synopses, data compression have been proposed as preprocessing tool for data stream mining.

Among dimensionality reduction techniques for data streams, time series representation criteria can be effectively used if the data stream paradigm is satisfied in the computational process. This is because real valued, append only data streams can be considered as continuously arriving time series. Usually these data are called "Streaming Time Series".

With this premise, we introduce a new strategy which provides a discretized representation of the incoming data by means of a string of symbols. It uses the knowledge extracted from a training set in order to get a more compact representation maintaining, at the same time, the main features of data as well as the same metric schema.

Moreover, we show how our strategy can be extended to represent multivariate streaming time series.

2 Related Works

Recently, there has been an increasing interest in developing strategies for time series dimensionality reduction in the data stream framework.

Some of these refer to some adjustments of techniques for stocked data while others have been proposed just for data streams. Among the existing techniques, we can mention: Discrete Fourier Transform (DFT) (Faloutsos et al. 1994), Discrete Wavelet Transform (DWT) (Kin-pong Chan and Ada Wai-Chee Fu 1999), Piecewise Linear Representation (PLR) (Morinaka et al. 2001), Piecewise Aggregate Approximation (PAA) (Yi and Faloutsos 2000), Symbolic Aggregate approximation (SAX) (Lin et al. 2003).

The main proposals based on DWT and DFT use sliding windows which move one step at a time, to update the transformation continuously over time. The few first coefficients are used for representing the streaming time series into a reduced space.

PLR consists in approximating data with a set of line segments. Since the number of segments is usually smaller than the time series size, a dimensionality reduction is so, performed.

The main criteria for the segments choice are linear interpolation and linear regression. The former is based on connecting the first and the last point of a subsequence, the latter uses the best fit segment in the sense of least squares.

In PAA the dimensionality reduction of the time series is performed splitting the incoming data into subsequences of equal-size and representing each one through its average value.

Starting from PAA, by means of SAX, a time series is transformed into a symbols sequence that can be processed using discrete data tools.

The basic idea is to map the PAA coefficients to symbols. The mapping rule is based on getting equi-probable regions from a normalized Gaussian distribution and to assign a symbol to each region. In such a way when a new batch of data arrives, the average value is computed and a symbol is assigned according to the region which includes such average value.

3 A New Symbolic Strategy for Streaming Time Series Dimensionality Reduction

Among time series reduction strategies, SAX is a very interesting one, both for its easy applicability and for its performance in several applicative fields. However, SAX is based on some assumptions which are often not met.

It gets equi-probable regions by means of a normalized Gaussian distribution, on the basis of the assumption that normalized time series have a Gaussian distribution. As shown in Fig. 1 this is not always true. Moreover if data are high frequency, an approach only based on the average value of windowed data could be not able to catch the true information if not by using a lot of symbols to make the string.

The here proposed approach, deals with these problems providing an effective solution to the time series representation challenge.

Given a time series $\mathbf{Y} = [(y_1, t_1), (y_2, t_2), \ldots, (y_i, t_i), \ldots, (y_N, t_N)]$ where $y_i \in \Re$ is a data point and t_i a time stamp, we represent $\mathbf{Y}$ through a string $\mathbf{S}$ of symbols.

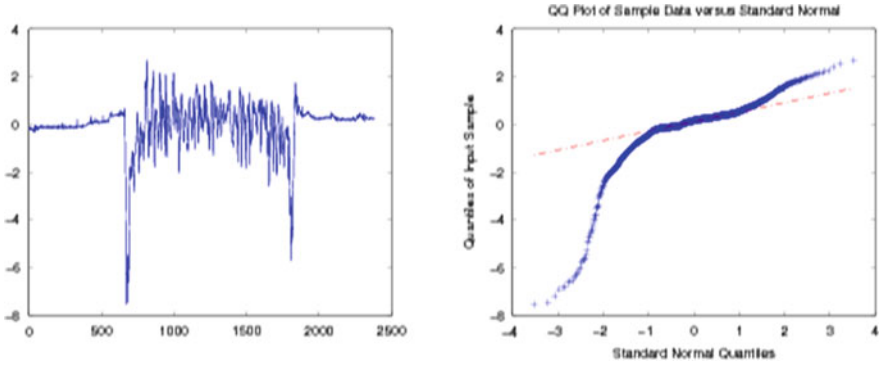


Fig. 1 Sensor signal - QQ plot

We propose to partition $\mathbf{Y}$ into W disjoint windows of equal size w_s and to map the data of each window to a couple of symbols. Each couple of symbols includes a "level" component and a "shape" component, in order to catch the shape of data in addition to their amplitude level.

In particular, starting from a streaming time series $\mathbf{Y}$, we get a symbolic string such as "1A 5F 7B 7A 4A 3C" where the numeric symbols summarizes the amplitude of data while the alphabetic symbol represents the shape underlying the data of each window.

The proposed approach shares some ideas with Vector Quantization (VQ) techniques which aim at finding the best approximation of a d-dimensional random vector X with distribution P by a random vector Y with at most n values in its image. As in VQ, we use the concept of alphabet (codebook) for approximating data, however we introduce specific features to facilitate data mining and on-line applications.

The strategy includes a training step, where a levels alphabet and a shapes alphabet are built on a training dataset, and a representation step of the on-line arriving data, according to the detected alphabets.

The training step has been introduced to take advantage of some apriori knowledge about data which is usually available also, in a streaming context. Moreover this approach provides a solution to the problem of using a Gaussian distribution to get the symbols regions used to map the data.

3.1 Training Step

In order to detect the two alphabets we split the data of the training dataset into windows of equal size w_s.

For the Levels alphabet, we have to build a rule to assign the level symbols to equi-probable regions of the time series domain.

This is performed by choosing the Levels alphabet size J and, then, by computing the histogram of the PAA coefficients in the training data set where the number of buckets is J.

This is to get a set $B = [b_1, b_2, \ldots, b_j, \ldots, b_J]$ (where $j \in J$) of equi-probable regions. Data are, then, mapped to the level symbols $L = [l_1, l_2, \ldots, l_j, \ldots, l_J]$ through the rule $b_j \Leftarrow l_j$.

To detect the Shape alphabet, we need a rule to map symbols to predefined data shapes.

To reach this aim, the data of each window are scaled by subtracting their average value and then processed by a clustering algorithm.

We propose to use a Dynamic Clustering Algorithm (DCA) (Diday 1971) since it looks both for a representation of the clusters and the best partition according to the minimization of a criterion function which is based on a suitable a dissimilarity measure.

The DCA performs a step of representation of the clusters and a step of allocation of the data according to the minimization of the distance of the data to the representative elements of the clusters (prototypes).

The clustering algorithm outputs a set $P = [p_1, p_2, \ldots, p_k, \ldots, p_K]$ (where $k \in K$ and K is the size of the shape alphabet) of representative elements (prototypes) of the clusters. These are mapped to the shape alphabet $S = [s_1, s_2, \ldots, s_k, \ldots, s_K]$ by the rule $p_k \Leftarrow s_k$, where s_k is a shape symbol.

3.2 Online Representation

Starting from the alphabets it is possible to compute the on-line representation of the streaming time series. The strategy can be summarized as follows:

- For each window
  1. Compute the average value $\bar{y_w}$
  2. Detect the level symbol for $\bar{y_w}$
  3. Detect the shape symbol comparing the shifted window data to the prototypes p_k with $k = 1, \ldots, K$
- End for each

When a new batch of data arrives, the algorithm searches for a suitable couple of symbols. The level symbol l_j is chosen by computing the average value $\bar{y_w}$ of data in a window and then detecting the symbol mapped to the histogram region which includes $\bar{y_w}$ (Fig. 2). To detect the shape symbol, we propose to shift the data by subtracting the average value $\bar{y_w}$ and then to find the symbol s_k mapped to the prototype p_k which minimizes the Euclidean distance between the shifted data and the prototypes.

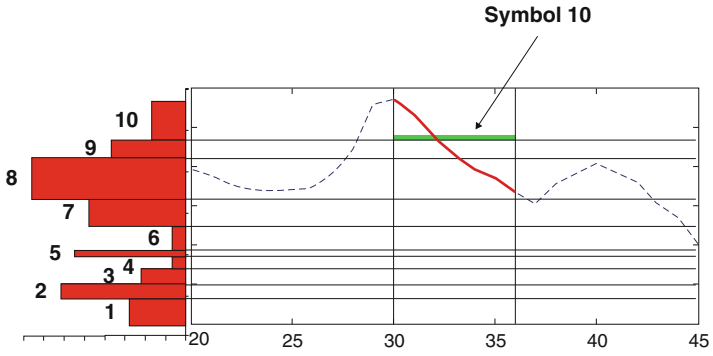


Fig. 2 Online detection of the Level symbol

3.3 A Feasible Representation for Bivariate Streaming Time Series

In the previous sections we have introduced our strategy for representing univariate time series. Here we extend such strategy for representing bivariate time series.

The main aim is to represent a bivariate time series by means of a string of couples of symbols just like in the univariate case.

Note that the here advised variations on the strategy proposed for univariate time series, are suitable for multivariate time series, however in order to maintain the effectiveness of the strategy, the alphabets size considerably increases.

The first adjustment is in the choice of the two alphabets.

In order to get the levels symbols, an algorithm for getting multivariate histograms is needed. An interesting proposal is Iacono and Irpino (2008) which quickly computes equi-probable rectangles from data. Each rectangle is here mapped to a level symbol l_j.

For the shape alphabet, the clustering algorithm used in the univariate case, is performed on the multidimensional trajectories, so to get multivariate prototypes.

The clusters prototypes are mapped to the shape symbols just like before, to form the shapes alphabet.

In the on-line algorithm, the step 1 has to be modified by introducing the computation of the average value for each variable in the window; in the step 3, the detection of the shape symbol is performed after shifting the single variables by their average value.

3.4 Time Series Approximation

From the symbolic string it is possible to build an approximation of the starting time series. The procedure can be summarized as follows:

- For each couple of symbols $[l_j s_k]$
  - Compute the average value of the region $\bar{b_j}$ mapped to l_j
  - Detect the shape p_k mapped to the current shape symbol s_k
  - Sum the average value $\bar{b_j}$ to the shape data p_k
- End for each

The procedure builds the approximated time series by joining the local representations detected from each couple of symbols. This is incrementally performed, at first, computing the average value of the region mapped to the level symbol and then summing the prototype mapped to the shape symbol.

4 Experimental Evaluation

In order to evaluate the effectiveness of the proposed strategy, we have performed tests on real data coming from several applicative contexts. Here we show the main results for one of these.

It is a dataset storing the electrical power consumptions of an home user. The recordings are performed, by means of a smart meter, every two minutes. The whole dataset includes 2 years of recordings.

The proposed strategy has been applied to this context, to evaluate how well it is able to approximate the collected data in term of mean square error.

Moreover, in order to compare our symbolic representation to the existing techniques, we have applied the PAA on the same dataset and then we have compared the approximation quality.

The choice of using the PAA has been performed since it shares with us the objective of producing an approximation of the represented time series to be used in data mining processes. This is not straightly included in SAX where the symbols summarize data through bounds instead of single identifiable values.

The first test consists in evaluating the impact of the windows size on the representation quality. The training set is made by power consumptions recorded in a week. To detect the shape alphabets, the well known k-means has been used since it is a particular case of DCA where the prototypes are the barycenter and the dissimilarity is the Euclidean distance.

The chosen alphabets size are: $L = 20$, $S = 8$. In order to make the comparison more effective, the windows size is chosen to be the double of the one chosen for PAA. This is to keep the same compression ratio since our strategy requires two coefficients for each window while PAA requires only one (However, note that PAA produces real valued approximations of the window data while the proposed strategy only uses a discrete set of values which require lesser storage resources).

Figure 3 shows how the means square error evolves by changing the window size for both the strategies. It is possible to note that our representation produces, for each window size, a better approximation quality compared to PAA.

A further test has been performed to evaluate the impact of the alphabets size on the quality of approximation. We have at first evaluated the MSE for several size of the Shape alphabet S setting $w_s = 35$ and $L = 20$ and then we have measured the influence of the Level alphabet size by setting $w_s = 35$ and $S = 6$. The results available in Fig. 4 highlight how both the two components the representation, affect the approximation quality.

5 Conclusions and Perspectives

In this paper we have proposed a new approach for representing time series by means of a string of symbols. The choice to use two alphabets facilitates knowledge discovery processes such as frequent pattern mining or clustering, since it allows to keep distinct the shape component from the amplitude component of data.

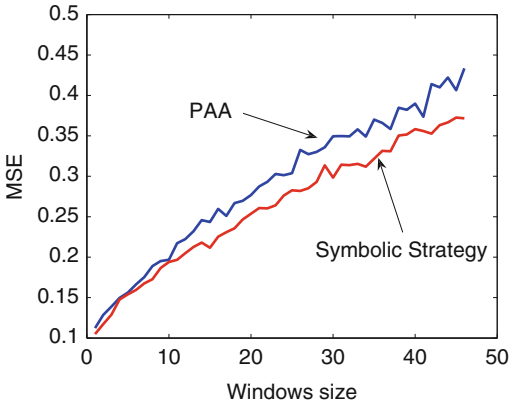


Fig. 3 Mean square error evaluation for different windows size

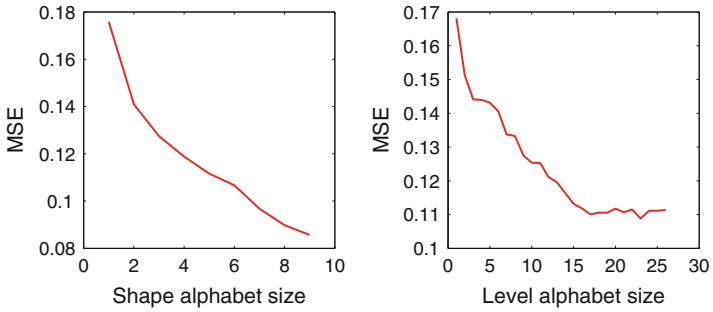


Fig. 4 Mean squared error evaluation for different alphabets size

However, the preliminary choice of the levels and shapes alphabets can introduce approximation problems when the knowledge about data is limited. This is especially true if there is an unpredictable change in the unknown statistical distribution underlying the data. To deal with this issue, further developments will be to update the alphabets when the prototypes and levels are no more able to ensure an high quality approximation of the original time series.

References

Diday, E. (1971). La Method des nuées dinamiques. In *Revue de Statistiques Appliquees XXX, 2*, 19–34.

Faloutsos, C., Ranganathan, M., & Manolopoulos, Y. (1994). Fast subsequence matching in time-series databases. In *Proceedings of the ACM SIGMOD Conference*. Minneapolis.

Iacono, M., & Irpino, A. (2008). Improving the MHIST-p algorithm for multivariate histograms of continuous data. In *Book of short papers of SFC-CLADAG*. Edizioni Scientifiche Italiane (ITALY). ISBN: 978-88-495-1656-2.

Kin-pong Chan, & Ada Wai-Chee Fu. (1999). Efficient time series matching by wavelets. *In Proceedings of the 15th International Conference on Data Engineering.*

Lin, J., Keogh, E., Lonardi, S., & Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. *In proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery.* June 13, San Diego, CA.

Morinaka, Y., Yoshikawa, M., Amagasa, T., & Uemura, S. (2001). The L-index: An indexing structure for efficient subsequence matching in time sequence databases. *In PAKDD* (pp. 51–60). Hong Kong.

Yi, B., & Faloutsos, C. (2000). Fast time sequence indexing for arbitrary lp norms. *In Proceedings of the 26th International Conference on Very Large Databases* (pp 385–394). Cairo, Egypt, Sept 10–14.

A Batesian Semiparametric Generalized Linear Model with Random Effects Using Dirichlet Process Priors

Kei Miyazaki and Kazuo Shigemasu

Abstract The purpose of this paper is to propose a parameter estimation method that doesn't need to set the number of heterogeneous populations in generalized linear models. We use a finite dimensional Dirichlet process mixed model (Ishwaran and James 2001). Due to the use of Dirichlet process, we make no assumption about the number of subgroups that are mixed. The proposed model can be considered as the direct extension of the model of Lenk and DeSarbo (2000) in the sense that the proposed method needs no assumption for the number of mixed latent classes in their model.

1 Introduction

The purpose of our presentation is to propose a parameter estimation method that doesn't need to set the number of heterogeneous populations in generalized linear models using Dirichlet process priors.

Generalized linear models (GLMs) can be applied to the data that follow the distributions of exponential family (McCullagh and Nelder 1983). Especially in behavioral sciences, because there are hardly any situations where the data are obtained from simply one population, finite mixture models that assume heterogeneous populations (latent classes) are often used (Wedel and DeSarbo 1995). Recently, finite mixture models that assume heterogeneous subpopulations (latent classes) and explain within-class heterogeneity by introducing random effects have been proposed (Lenk and DeSarbo 2000).

The existing methods require that the number of subgroups is determined in advance and moreover, these methods require the calculation of information criterion in order to determine the number of subgroups, which usually results in a heavy computational burden. The method could be useful that made it possible to estimate the parameters and the number of latent classes simultaneously.

K. Miyazaki (✉)
Department of Cognitive and Behavioral Science, The University of Tokyo, Komaba 3-8-1, Meguro-ku, Tokyo, 153-8902 Japan
e-mail: miyazaki.behaviormetrics@gmail.com

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_42,

In this presentation, we use a Batesian estimation method with a finite Dirichlet process mixed model (Ishwaran and James 2001). Due to the use of a Dirichlet process, we make no assumption about the number of subgroups that are mixed. Kleinman and Ibrahim (1998) introduce Dirichlet process as a method that can assume any forms of distribution. Whereas, we introduce Dirichlet process as a method that can make it unnecessary to set the number of subgroups (latent class) before parameter estimation.

The proposed model can also be considered as the direct extension of the model of Lenk and DeSarbo (2000) in the sense that the proposed method needs no assumption for the number of mixed latent classes in their model.

2 Finite Mixture GLM with Random Effects

We begin with reviewing the finite mixture GLM with random effects (Lenk and DeSarbo 2000). There are several observations on each subject. i-th subject has J observations. Let y_{ij} be the j-th dependent observation on subject i and $\boldsymbol{x}_{ij}$ be the corresponding $p \times 1$ vector of independent variables. Define:

$$\boldsymbol{y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iJ} \end{pmatrix} \quad X_i = \begin{pmatrix} \boldsymbol{x}'_{i1} \\ \vdots \\ \boldsymbol{x}'_{iJ} \end{pmatrix} \quad \text{for } i = 1, \ldots, n \tag{1}$$

The probability density function of y_{ij} is expressed as follows:

$$f(y_{ij}|\boldsymbol{\beta}_i) = \exp\left[\frac{y_{ij}h(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_i) - b[h(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_i)]}{a(\phi_i)} + c(y_{ij}, \phi_i)\right] \tag{2}$$

ϕ_i is the scale parameter and a, b, c, h are functions depending on the member of the exponential family. We use the expression $k_i = l$ to imply that the i-th examinee belongs to the l-th component. When $k_i = l$, $\boldsymbol{\beta}_i (p \times 1)$ follows the normal distribution with mean vector of $\boldsymbol{\theta}_l = (\theta_{1l}, \ldots, \theta_{pl})'$ and variance-covariance matrix of Ψ_l

$$\boldsymbol{\beta}_i | k_i = l \ \sim \ N(\boldsymbol{\theta}_l, \boldsymbol{\Psi}_l) \tag{3}$$

3 Representation of the Dirichlet Process Mixture Model

According to Sethuraman (1994), when Dirichlet process priors $F \sim DP(\alpha, G_0)$ are assumed, F is expressed as follows:

$$F(\cdot) = \sum_{l=1}^{\infty} \kappa_l \delta_{\xi_l}(\cdot), \quad \xi_l \sim G_0 \tag{4}$$

where δ_{ξ_l} denotes a discrete measure concentrated at ξ_l, $\kappa_l = \prod_{k=1}^{l-1}(1 - V_k)V_l$ and $V_1, V_2, \ldots$ independently follow a beta distribution $\mathrm{Be}(1, \alpha)$. In this, α is the parameter that indicates ease of transition to other components while G_0 indicates a reference distribution. Parameters are generated from this distribution when the new component is generated. When the above Dirichlet process priors are assumed, a random variable vector $\boldsymbol{y}$ is expressed by the following Dirichlet process mixture models:

$$\boldsymbol{y} \sim \sum_{l=1}^{\infty} \kappa_l f(\cdot|\boldsymbol{\psi}_l) \tag{5}$$

where f is the sampling distribution of data $\boldsymbol{y}$ and $\boldsymbol{\psi}_l$ is a parameter vector (refer to Walker et al. 1999 for a detailed explanation of the Dirichlet process mixture model). This equation indicates that any distribution can be expressed as a mixture distribution of conventional distributions such as normal distributions, and it is not necessary to set the number of mixed components for analysis.

Ishwaran and Zarepour (2000) proposed the following finite-dimensional Dirichlet process priors:

$$F(\cdot) = \sum_{l=1}^{L} \kappa_l \delta_{\xi_l}(\cdot), \quad \xi_l \sim G_0 \tag{6}$$

Assuming the above finite-dimensional Dirichlet process priors, a random variable vector $\boldsymbol{y}$ is expressed by the following finite-dimensional Dirichlet process mixture models:

$$\boldsymbol{y} \sim \sum_{l=1}^{L} \kappa_l f(\cdot|\boldsymbol{\psi}_l) \tag{7}$$

where L is the maximum number of components. Ishwaran and James (2001) (Theorem 2) proved that (7) approximates infinite-dimensional Dirichlet process mixture models with satisfactory accuracy when the value of L is large enough and the value of the upper bound of errors that changes according to the sample size and the value of L is described with sketches of proofs. According to their paper, the truncation value of L has more influence than the sample size on reducing the error (Theorem 2).

Prior Distributions

The prior distributions for the mean vector $\boldsymbol{\theta}$ and variance-covariance matrix $\boldsymbol{\Psi}$ are the multivariate normals, inversed Wishart distributions, as follows:

$$\begin{aligned} \boldsymbol{\theta} &\sim DP_L(\alpha, N(\boldsymbol{\mu}_0, \Sigma_0)) \\ \boldsymbol{\Psi} &\sim DP_L(\alpha, IW(\nu_0, S_0) \end{aligned} \tag{8}$$

DP_L denotes L-dimensional finite Dirichlet process prior and IW the inverted Wishart distribution. The values of hyper parameters $\boldsymbol{\mu}_0, \Sigma_0, \nu_0, S_0$ are fixed (see Sect. 5 simulation studies). α is the parameter that indicates ease of transition to other components.

4 Algorithm: Blocked Gibbs Sampler

In this section, we apply the Blocked Gibbs Sampler proposed by Ishwaran and James (2001) to our semiparametric generalized linear model. We describe the Blocked Gibbs Sampler for our model setup and introduce the full conditional distributions of each parameter's vector necessary to draw samples in each iteration. Refer to Diebolt and Robert (1994) who recommended the Markov Chain Monte Carlo algorithms for finite mixture models. Our algorithm is based on the works of Ishwaran and James (2001) as well as Diebolt and Robert (1994).

Algorithm and the Full Conditional Distributions
Let $\{k_1^*, \ldots, k_m^*\}$ be the set of the current m unique values of $\boldsymbol{k}$. '$\cdots$' means that the other parameters are given. Then to run the Blocked Gibbs Sampler, we draw parameter values in the following order:

(1) $\boldsymbol{\beta}$: Generate $\boldsymbol{\beta}$ from the following full-conditional distribution:

$$p(\boldsymbol{\beta}_i|\cdots) \propto p(\boldsymbol{y}_i|\boldsymbol{\beta}_i)p(\boldsymbol{\beta}_i|k_i = l, \boldsymbol{\theta}_l, \boldsymbol{\Psi}_l)$$
$$\propto \exp\left[\frac{y_{ij}h(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_i) - b[h(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_i)]}{a(\phi_i)} - \frac{1}{2}(\boldsymbol{\beta}_i - \boldsymbol{\theta}_l)'\boldsymbol{\Psi}_l^{-1}(\boldsymbol{\beta}_i - \boldsymbol{\theta}_l)\right] \tag{9}$$

$\boldsymbol{\beta}_i$ is generated with the aid of a Metropolis step. Refer to Lenk and DeSarbo (2000) who recommended the method of generating β. This step is based on the work of them.

(2) $\boldsymbol{\theta}$: Generate $\boldsymbol{\theta}_l$ from $p(\boldsymbol{\theta}_l|\boldsymbol{\tau}_\theta)$ for each $k \in \boldsymbol{k} - \{k_1^*, \ldots, k_m^*\}$ (for components to which no subject has been allocated). For components to which at least one subject has been allocated, $\boldsymbol{\theta}_l$ is generated from the following full-conditional distribution

$$p(\boldsymbol{\theta}_l|\cdots) \propto \prod_{i:k_i=l} p(\boldsymbol{\beta}_i|k_i = l, \boldsymbol{\theta}_l, \boldsymbol{\Psi}_l)p(\boldsymbol{\theta}_l|\boldsymbol{\tau}_\theta)$$
$$\propto \exp\left[-\frac{1}{2}(\boldsymbol{\theta}_l - \boldsymbol{\mu}_l)'\Sigma_l^{-1}(\boldsymbol{\theta}_l - \boldsymbol{\mu}_l)\right] \tag{10}$$
$$\Sigma_l = (n_l\boldsymbol{\Psi}_l^{-1} + \Sigma_{0l}^{-1})^{-1},\ \boldsymbol{\mu}_l = \Sigma_l(n_l\boldsymbol{\Psi}_l^{-1}\bar{\boldsymbol{\beta}}_l + \Sigma_{0l}^{-1}\boldsymbol{\mu}_{0l})$$
$$n_l = \sum_{i:k_i=l} 1,\ \bar{\boldsymbol{\beta}}_l = n_l^{-1}\sum_{i:k_i=l}\boldsymbol{\beta}_i$$

(3) $\boldsymbol{\Psi}$: Generate $\boldsymbol{\Psi}_l$ from $p(\boldsymbol{\Psi}_l|\boldsymbol{\tau}_\Psi)$ for each $k \in \boldsymbol{k} - \{k_1^*, \ldots, k_m^*\}$ (for components to which no subject has been allocated). For components to which at least one subject has been allocated, $\boldsymbol{\Psi}_l$ is generated from the following full-conditional distribution

$$
\begin{aligned}
p(\boldsymbol{\Psi}_l|\cdots) &\propto \prod_{i:k_i=l} p(\boldsymbol{\beta}_i|k_i=l,\boldsymbol{\theta}_l,\boldsymbol{\Psi}_l)\,p(\boldsymbol{\Psi}_l|\boldsymbol{\tau}_\Psi) \\
&\propto |\boldsymbol{\Psi}_l|^{-(\nu_l+p+1)/2}\exp\left[-\frac{1}{2}\mathrm{tr}\{\boldsymbol{\Psi}_l^{-1}S_l\}\right] \qquad (11)\\
\nu_l &= \nu_{0l}+n_l,\ S_l = S_{0l}+\sum_{i:k_i=l}(\boldsymbol{\beta}_i-\boldsymbol{\theta}_l)(\boldsymbol{\beta}_i-\boldsymbol{\theta}_l)'
\end{aligned}
$$

(4) $\boldsymbol{k}$: Generate k_i from the following distribution

$$
p(k_i|\cdots) \sim \sum_{l=1}^{L}\pi_{li}\delta_l(\cdot) \qquad (12)
$$

where

$$
\pi_{li} = \frac{\kappa_l\, p(\boldsymbol{\beta}_i|k_i=l,\boldsymbol{\theta}_l,\boldsymbol{\Psi}_l)}{\sum_{l=1}^{L}\kappa_l\, p(\boldsymbol{\beta}_i|k_i=l,\boldsymbol{\theta}_l,\boldsymbol{\Psi}_l)} \qquad (13)
$$

(5) $\boldsymbol{\kappa}$: the full conditional distribution of $\boldsymbol{\kappa}$ is the generalised Dirichlet distribution:

$$
\begin{aligned}
\kappa_l &= \prod_{m=1}^{l-1}(1-V_m)V_l \\
V_l &\sim \mathrm{Be}(a_l+M_l, b_l+\sum_{m=l+1}^{L}M_m)
\end{aligned} \qquad (14)
$$

and M_l is the number of k_i that equals l.

By using a Dirichlet process, we can evaluate the posterior probabilities of the numbers of components drawn by this sampler.

5 Simulation Studies

In order to prove the reliability of the proposed method, we conducted two simulation studies. Via two simulation studies, we will verify that for a known number of mixture components the MCMC procedure recovers the unknown parameters.

Simulation Study 1
In the first simulation study we generated 100 data sets from a linear regression model with normal error, and for each data set, we obtained the usual Bayesian estimates as well as estimates using the proposed method. We were interested in how accurately our method estimated the parameters. The number of the dependent and the explanatory variables was 10 and 3, respectively. The number of true components was 2. Keeping in mind the upper bound of errors caused by using

Table 1 The true and estimated values of parameters in simulation study 1

	True values	Estimates	RMS		True values	Estimates	RMS
κ_1	0.7	0.688	4.70×10^{-2}	Ψ_{111}	1.0	0.977	0.135
κ_2	0.3	0.296	3.78×10^{-2}	$\Psi_{211}(=\Psi_{121})$	−0.5	−0.489	0.137
κ_3	*	9.94×10^{-3}	2.75×10^{-2}	$\Psi_{311}(=\Psi_{131})$	1.5	1.47	0.208
κ_4	*	3.17×10^{-3}	4.00×10^{-3}	Ψ_{221}	2.0	1.99	0.234
κ_5	*	1.59×10^{-3}	1.70×10^{-3}	$\Psi_{321}(=\Psi_{231})$	−1.0	−0.989	0.238
κ_6	*	8.27×10^{-4}	8.35×10^{-4}	Ψ_{331}	3.0	2.95	0.413
κ_7	*	4.15×10^{-4}	4.41×10^{-4}	Ψ_{112}	3.0	2.95	0.661
κ_8	*	1.87×10^{-4}	1.89×10^{-4}	$\Psi_{212}(=\Psi_{122})$	0.5	0.565	0.326
κ_9	*	6.85×10^{-5}	6.92×10^{-5}	$\Psi_{312}(=\Psi_{132})$	−0.8	−0.749	0.438
κ_{10}	*	1.48×10^{-5}	1.49×10^{-5}	Ψ_{222}	1.0	1.07	0.255
θ_{11}	0	6.78×10^{-3}	7.83×10^{-2}	$\Psi_{322}(=\Psi_{232})$	0.3	0.279	0.237
θ_{21}	2.0	1.97	0.134	Ψ_{332}	2.0	1.97	0.521
θ_{31}	−2.0	−1.98	0.147	*	*	*	*
θ_{12}	−2.0	−1.98	0.301	*	*	*	*
θ_{22}	0	3.46×10^{-2}	0.176	*	*	*	*
θ_{32}	2.0	2.03	0.322	*	*	*	*

Table 2 The true and estimated values of parameters in simulation study 2

	True values	Estimates	RMS		True values	Estimates	RMS
κ_1	0.6	0.516	0.118	θ_{11}	0	−0.0755	0.275
κ_2	0.4	0.306	0.111	θ_{21}	−1.0	−0.817	0.548
κ_3	*	0.102	0.122	θ_{12}	−1.0	−0.873	0.333
κ_4	*	0.0467	0.0625	θ_{22}	1.0	0.737	0.706
κ_5	*	0.0184	0.0256	Ψ_{111}	0.02	0.0219	0.0151
κ_6	*	7.25×10^{-3}	0.0119	$\Psi_{211}(=\Psi_{121})$	0.03	0.0253	0.0179
κ_7	*	2.63×10^{-3}	5.31×10^{-3}	Ψ_{221}	0.07	0.0527	0.0255
κ_8	*	5.59×10^{-4}	1.34×10^{-3}	Ψ_{112}	0.07	0.0542	0.0337
κ_9	*	1.10×10^{-4}	4.06×10^{-4}	$\Psi_{212}(=\Psi_{122})$	−0.03	−0.0165	0.0244
κ_{10}	*	1.19×10^{-5}	1.92×10^{-5}	Ψ_{222}	0.02	0.0143	0.0177

finite-dimensional Dirichlet process priors, as shown by Ishwaran and James (2001), we set γ and L as 3 and 10, respectively. We fixed the hyperparameters as follows: $\boldsymbol{\mu}_{0l} = \mathbf{0}, \Sigma_{0l} = 100I, \nu_{0l} = p + 1 = 4, S_{0l} = 0.1I$ (common across components). For every 100 trials, we generated 200 observations, all of which followed the same population parameters (or true values) as given in Table 1. After employing 2,000 burn-in iterations, we employed 3,000 Gibbs iterations to calculate the numerical posterior distributions or the posterior moments. For each of the 100 data sets, we calculated the parameters estimates and averaged the 100 sets of estimates, and in order to check the accuracy of the results of this simulation, we calculated Root mean squares (RMS) between estimates and true values for each data sets and averaged them. These results are listed in Table 1. These results indicate that the proposed method yielded accurate estimates of the parameters.

5.1 Simulation Study 2

For the second simulation study, we generated 50 data sets from a logistic regerssion model. As in the first simulation study, we also set γ and L to 3 and 10, respectively, and set 3,000 observations for each data set. The number of true components was 2. After employing 2,000 burn-in iterations, we employed 3,000 Gibbs iterations to calculate the numerical posterior distributions or the posterior moments. The number of the dependent and the explanatory variables was 50 and 2, respectively. We fixed the hyperparameters as follows: $\boldsymbol{\mu}_{0l} = \mathbf{0}, \Sigma_{0l} = 100000I, \nu_{0l} = p + 1 = 3, S_{0l} = 0.00001I$ (common across components). The true parameters and the quality of their estimates can be assessed by Table 2. As criterions to compare the accuracy of the results of this simulation, for each data set we generated 1,000 observed values of dependent variables from the true model. Thereafter, we calculated predictive values of dependent variables using the proposed model, single component GLM (that is, standard GLM), respectively. We calculated the Root Mean Squared errors between the true values of dependent values and the estimates of them. The obtained values were 0.619 and 2.12 for our proposed model and a standard model, respectively. RMSs were approximately 3.4 times larger than that of the proposed method. This result therefore also indicates that the proposed method yields accurate estimates of the parameters, whereas the standard model essentially yields biased estimates.

6 Conclusion

In this study, we proposed a semiparametric Bayesian estimation method in generalized linear models using Dirichlet process priors. As we know from the study of Sethuraman (1994), any shapes of distributions are expressed through a mixture of probability density functions of conventional distributions such as normal distributions. Applying Dirichlet process mixture models that do not rely on an assumption of the number of mixed components, our new semiparametric generalized linear model can estimate both the parameters as well as the number of mixed components.

The existing methods cannot simultaneously estimate the parameters and determine the number of components. The analysis process using the existing methods is as follows: when Bayesian estimation is used on 1 to some number of components, construct some number of models, estimate the parameters for each model and find the most appropriate model using criteria such as AIC, BIC or marginal likelihood. These criteria are usually difficult to calculate. To complete one study, one would need to construct and run at least two kinds of programs. On the other hand, our method does not require that the most appropriate number of components to be mixed be decided. In other words, when using this method, one only has to construct and run one kind of program, giving our method an advantage over the existing methods.

Dirichlet process mixture modelling is often used to construct semiparametric or nonparametric Bayesian modelling, and our model is largely consistent with this

idea. In other words, it can be said that we have proposed a new semiparametric Bayesian generalized linear model. Owing to the use of Bayesian estimation, our method is effective even when the sample size is small.

Acknowledgements This study was supported by Grant-in-Aid for Scientific Research, 19-8879. We would like to express our sincere thanks to the editor and the reviewer for their valuable advice and comments.

References

Diebolt, J., & Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B, 56*, 363–375.

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for Stick-Breaking priors. *Journal of the American Statistical Association, 96*, 161–173.

Ishwaran, H., & Zarepour, M. (2000). Markov Chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika, 87*, 371–390.

Kleinman, K. P., & Ibrahim, J. G. (1998). A semi-parametric bayesian approach to generalized linear mixed models. *Statistics in Medicine, 17*, 2579–2596.

Lenk, P. J., & DeSarbo, W. S. (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika, 65*, 93–119.

McCullagh, P., & Nelder, J. A. (1983). *Generalized linear models*. New York: Chapman and Hall.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica, 4*, 639–650.

Wedel, M., & DeSarbo, W. S. (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification, 12*, 21–55.

Walker, S. G., Damien, P., Laud, P. W., & Smith, A. F. M. (1999). Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society, Series B, 61*, 485–527.

Exact Confidence Intervals for Odds Ratios with Algebraic Statistics

Anne Krampe and Sonja Kuhnt

Abstract Odds ratios, which compare the odds of an event occurring in the presence of a potential risk-factor to the odds of it occurring in the absence of the potential risk-factor, are commonly used in medical and social science research. Confidence intervals usually rely on approximate results or exact enumeration. We suggest an algebraic solution to this problem which is of particular use in situations where the approximation is not adequate and exact enumerations are computationally too costly. This algebraic approach relies on the Diaconis-Sturmfels algorithm which combines computational commutative algebra and Markov chain Monte Carlo methods to simulate samples of the conditional distribution of a discrete exponential family with given sufficient statistic. In particular a Groebner basis is used for the construction of the Markov chain. In a simulation study we determine and compare the simulated, exact and approximate results.

1 Introduction

Categorical data occur in various statistical applications. In many situations it is not only of interest to detect a dependency between two variables but rather to estimate the strength of the relationship. In clinical studies or epidemiology, for example, it is often of interest to identify so-called risk-factors. These factors increase the risk of contracting a disease. The odds ratio (OR) compares the chance to fall ill under exposure with the chance to fall ill under no exposure. Formally, we consider two binary random variables X and Y with outcome $i \in \{1,2\}$ and $j \in \{1,2\}$. The odds ratio is defined as cross-product ratio $OR = \frac{P(Y=1|X=1)/P(Y=2|X=1)}{P(Y=1|X=2)/P(Y=2|X=2)}$. Let n_{ij} denote the counts in a sample of size n and N_{ij} the random variable. The Maximum-Likelihood estimator for the odds ratio is given by $\widehat{OR} = \frac{N_{11} \cdot N_{22}}{N_{12} \cdot N_{21}}$. Modified versions of the estimator have been proposed to account for the situation where n_{12} or n_{21}

S. Kuhnt (✉)
Faculty of Statistics, TU Dortmund University, Germany
e-mail: kuhnt@statistik.tu-dortmund.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_43,

are equal to zero (cf. Haldane 1955; Fleiss et al. 2003; Agresti 2002). Here, we stay with the common ML estimator. The resulting confidence interval allows us to draw conclusions concerning the statistical significance. Usually, the confidence interval is based on an approximation which might not be adequate, e.g. if the data are sparse. Mehta et al. (1985) suggest an exact confidence interval conform to Fisher's exact test for independence, which, however, is known to be conservative.

In this paper we develop an alternative confidence interval. In algebraic statistics results from algebraic geometry are used to address statistical problems (Riccomagno 2009). In a key paper, Diaconis and Sturmfels (1998) combine computational algebra and statistics via Markov chain Monte Carlo methods. Using the Metropolis-Hastings algorithm (Chib and Greenberg 1995; Sørensen and Gianola 2002), a Markov chain is generated whose stationary distribution equals the conditional distribution of a discrete exponential family with given sufficient statistic. We exploit these results to derive an algebraic confidence interval based on a simulated exact distribution of the estimated odds ratio.

2 Traditional Confidence Intervals for the Odds Ratio

For inference, it is convenient to consider the log-transformed odds ratio, $\log(OR)$. Woolf's approximate confidence interval for the odds ratio relies on the asymptotic normal distribution of $\log(\widehat{OR})$ with mean $\log(OR)$ (Woolf 1971). Using the delta-method it is easy to show that the variance of $\log(\widehat{OR})$ can be estimated by

$$\widehat{\sigma}^2(\log(\widehat{OR})) = \frac{1}{N_{11}} + \frac{1}{N_{12}} + \frac{1}{N_{21}} + \frac{1}{N_{22}}.$$

Hence, it holds that

$$P\left(-u_{1-\alpha/2} \leq \frac{\log(\widehat{OR}) - \log(OR)}{\widehat{\sigma}(\log(\widehat{OR}))} \leq u_{1-\alpha/2}\right) \approx 1-\alpha,$$

where $u_{1-\alpha/2}$ denotes the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution. After some calculation we get the $(1-\alpha)\cdot 100\%$-confidence interval

$$CI(OR) := \left[\widehat{OR} \cdot \exp\{\pm u_{1-\alpha/2} \cdot \widehat{\sigma}(\log(\widehat{OR}))\}\right].$$

Note that this confidence interval does not exist if the value of $\widehat{OR}$ is zero or ∞. In such cases the lower limit is set to zero and the upper limit of this confidence interval is set to ∞, respectively.

Using the idea of Fisher's exact test for independence we can construct an exact confidence interval for the odds ratio. We assume that N_{ij}, $i, j = 1, 2$, are multinomially distributed. Conditional on the observation of the sufficient statistic for the

parameters of the independence model, $t = (n_{1+}, n_{2+}, n_{+1}, n_{+2})'$, the distribution of N_{11} depends only on OR and is noncentral hypergeometric

$$f(N_{11} = x|\, t, OR) = \frac{\binom{n_{1+}}{x} \cdot \binom{n_{2+}}{n_{+1}-x} \cdot (OR)^x}{\sum_{u=m_-}^{m^+} \binom{n_{1+}}{u} \cdot \binom{n_{2+}}{n_{+1}-u} \cdot (OR)^u}, \quad m_- \le x \le m^+;$$

with $m_- = \max(0, n_{1+} + n_{+1} - n)$ and $m^+ = \min(n_{1+}, n_{+1})$, see Fleiss et al. (2003), Mehta et al. (1985), and Zelen (1971) for details. The lower limit OR_* of the exact $(1-\alpha)\cdot 100\%$-confidence interval for the odds ratio is determined as follows

$$OR_* = 0 \text{ if } n_{11} = m_-, \text{ and } \sum_{x=n_{11}}^{m^+} f(x|\, t, OR_*) = \frac{\alpha}{2} \text{ if } m_- < n_{11} \le m^+.$$

The upper exact limit OR^* fulfills the condition

$$\sum_{x=m_-}^{n_{11}} f(x|\, t, OR^*) = \frac{\alpha}{2} \text{ if } m_- \le n_{11} < m^+, \text{ and } OR^* = \infty \text{ if } n_{11} = m^+.$$

3 Algebraic Confidence Interval

We now develop a new confidence interval for the odds ratio as an alternative to the traditional methods. Its construction relies on both, the exact and the approximate confidence interval. In particular, we replace the normal approximation in the common confidence interval by a simulation of the exact one. The exact confidence interval relies on the noncentral hypergeometric distribution on the set of all data sets with observed sufficient statistic t, which we denote by $\mathcal{Z}_t$. This is a conditional distribution of a discrete exponential family given the sufficient statistic t. For this general situation Diaconis and Sturmfels (1998) suggest the concept of a so-called Markov basis to construct an appropriate proposal distribution for the Metropolis-Hastings algorithm. Thereby a Markov chain can be generated with the hypergeometric distribution as stationary distribution, and we are able to substitute the exact distribution by a simulation in cases where an exact enumeration becomes unfeasible and the approximate distribution might not yet be applicable.

To apply the Diaconis and Sturmfels algorithm (Diaconis and Sturmfels 1998) it is essential to display $\mathcal{Z}_t$ as

$$\mathcal{Z}_t := \{z : \mathcal{H} \to \mathbb{N} |\, \sum_{x \in \mathcal{H}} z(x) T^*(x) = t\},$$

where $\mathcal{H}$ is the finite sample space, T^* the adequate mapping from $\mathcal{H}$ to $\mathbb{N}^d$ and d the dimension of t. For 2×2 tables we have $x = (i, j) \in \mathcal{H} = \{(i, j) \mid i, j = 1, 2\}$. The joint probability function of the multinomially distributed N_{ij}, $i, j = 1, 2$, belongs to a four-parametrical exponential family and the sufficient statistic for the parameters in this independence model consists of the row and column sums of the given table, hence $T = (N_{1+}, N_{2+}, N_{+1}, N_{+2})'$. We can now identify T^* for a 2×2 table under the assumption of independence. $T^*((i, j))$ is a vector of the same length than the sufficient statistic with two entries equal to one at positions i and $2 + j$, the other entries are zero. Hence, we get

$$T^*((1,1)) = (1,0,1,0)' \qquad T^*((1,2)) = (1,0,0,1)'$$
$$T^*((2,1)) = (0,1,1,0)' \qquad T^*((2,2)) = (0,1,0,1)'.$$

A Markov basis is defined as a set of functions $m_1, m_2, \ldots, m_L : \mathcal{H} \to \mathbb{Z}$, called moves, such that

- $\sum_{x\in\mathcal{H}} m_i(x)T^*(x) = 0, \quad$ for all $1 \le i \le L \quad$ and
- for any t and $z, z' \in \mathcal{Z}_t$ there is a sequence of moves $(m_{i_1}, \ldots, m_{i_A})$ as well as a sequence of directions $(\epsilon_1, \ldots, \epsilon_A)$ with $\epsilon_j = \pm1$, such that

$$z' = z + \sum_{j=1}^{A} \epsilon_j m_{i_j} \quad \text{and} \quad z + \sum_{j=1}^{a} \epsilon_j m_{i_j} \ge 0, \quad forall\ 1 \le a \le A.$$

Hence, adding or subtracting sequences of moves to the observed table does not change the value of the sufficient statistic and each element of $\mathcal{Z}_t$ can be reached by such a sequence. Now computational algebra becomes relevant as each x in the sample space $\mathcal{H}$ is identified by an indeterminant, also denoted by x and we let $k[\mathcal{H}]$ be the ring of polynomials in these indeterminants. Further, any function $g : \mathcal{H} \to \mathbb{N}$ is identified by a monomial $\prod_{x\in\mathcal{H}} x^{g(x)}$. Now consider the auxiliary ideal $\mathcal{I}_a = \{x - \mathcal{T}^{T^*(x)},\ x \in \mathcal{H}\}$ with $\mathcal{T}^{T^*(x)} := T_1^{T_1^*(x)} \cdot T_2^{T_2^*(x)} \cdots T_d^{T_d^*(x)}$ and T_i^* the i^{th} component of T^* according to the considered model, see e.g. Krampe and Kuhnt (2009). Denote the reduced Gröbner basis of $\mathcal{I}_a$ by $\mathcal{G}_a$ and set $\mathcal{I}_T := \mathcal{I}_a \cap k[\mathcal{H}]$. The reduced Gröbner basis $\mathcal{G}$ of $\mathcal{I}_T$ consists of those polynomials of $\mathcal{G}_a$ that only involve elements of $\mathcal{H}$. Diaconis and Sturmfels (1998) (Theorems 3.1 and 3.2) show that $\mathcal{G}$ equals the Markov basis needed for the Metropolis-Hastings algorithm.

Here, we calculate the Gröbner basis $\mathcal{G}_a$ of the ideal $\mathcal{I}_a$, $\mathcal{I}_a =< x_{11} - N_{1+}N_{+1},\ x_{12} - N_{1+}N_{+2},\ x_{21} - N_{2+}N_{+1},\ x_{22} - N_{2+}N_{+2}, >$. Assuming graded lexicographical ordering the resulting Gröbner basis for the independence model for a 2×2 table is $\mathcal{G} = \{x_{11}x_{22} - x_{12}x_{21}\}$. This polynomial can be considered as a matrix. Note that each function $m : \mathcal{H} \to \mathbb{Z}$ can be rewritten as $m(x) = m^+(x) - m^-(x)$ with $m^+, m^- : \mathcal{H} \to \mathbb{N}$, $m^+(x) = \max(m(x), 0)$ and $m^-(x) = \max(-m(x), 0)$. We can thereby write the polynomial, $x_{11}^1 x_{12}^0 x_{21}^0 x_{22}^1 - x_{11}^0 x_{12}^1 x_{21}^1 x_{22}^0 = x_{11}x_{22} - x_{12}x_{21}$,

as table (or move)

$$\begin{pmatrix} +1 & -1 \\ -1 & +1 \end{pmatrix}.$$

Hence, the allowed move of the table used to generate data sets with the same observed sufficient statistic as the considered table is intuitive.

Based on the moves from the Markov basis the Markov chain using the Metropolis-Hastings algorithm is constructed as follows, cf. Rapallo (2003):

- Choose a move m_U uniformly on $\{1, \ldots, L\}$ and a direction of the move $\epsilon = \pm 1$ with probability $\frac{1}{2}$ independently of U.
- Assume that the chain is currently in state $z \in \mathcal{Z}_t$.
- The chain moves to $z' = z + \epsilon m_U \in \mathcal{Z}_t$ with probability

$$\alpha = \min\left(\frac{H(z')}{H(z)}, 1\right) = \min\left(\frac{\prod\limits_{x \in \mathcal{H}} z(x)!}{\prod\limits_{x \in \mathcal{H}} (z(x) + \epsilon m_U(x))!}, 1\right).$$

If the suggested new state z' has a negative entry, the hypergeometric density $H(z')$ and thereby α are zero and hence, the chain stays in $z \in \mathcal{Z}_t$.

Practically, we generate a Markov chain with l states. According to the usual MCMC approach we disregard the initial b states in the so-called burn-in-phase and then sample each sth table. For each of the remaining $\lfloor \frac{l-b}{s} \rfloor$ tables we calculate the logarithm of the estimated odds ratio $\log(\widehat{OR})$ and thereby determine the conditional distribution of $\log(\widehat{OR})$, assuming X_1 and X_2 are independent. By considering $\frac{\log(\widehat{OR}) - \log(OR)}{\sqrt{\widehat{\sigma}(\log(\widehat{OR}))}}$ we switch over to a standardized statistic. Note that $\log(OR)$ equals zero under the hypothesis of independence. According to Woolf's approximate confidence interval for the odds ratio, we use for the algebraic confidence interval that

$$P\left(q_{\text{alg},\,\frac{\alpha}{2}} \leq \frac{\log(\widehat{OR}) - \log(OR)}{\sqrt{\widehat{\sigma}(\log(\widehat{OR}))}} \leq q_{\text{alg},1-\frac{\alpha}{2}}\right) \approx 1 - \alpha,$$

where the simulated algebraic quantiles q_{alg} replace the approximate quantiles of the standard normal distribution. Hence, the algebraic $(1-\alpha) \cdot 100\%$ confidence interval for the odds ratio is given by

$$CI(OR) := \left[\widehat{OR} \cdot \exp\left\{-q_{\text{alg},\,1-\frac{\alpha}{2}} \cdot \sqrt{\widehat{\sigma}(\log(\widehat{OR}))}\right\};\right.$$
$$\left.\widehat{OR} \cdot \exp\left\{-q_{\text{alg},\,\frac{\alpha}{2}} \cdot \sqrt{\widehat{\sigma}(\log(\widehat{OR}))}\right\}\right].$$

Table 1 Cell probabilities for the simulation model with $OR = 1$

$$\begin{bmatrix} 0.278 & 0.278 \\ 0.222 & 0.222 \end{bmatrix}$$

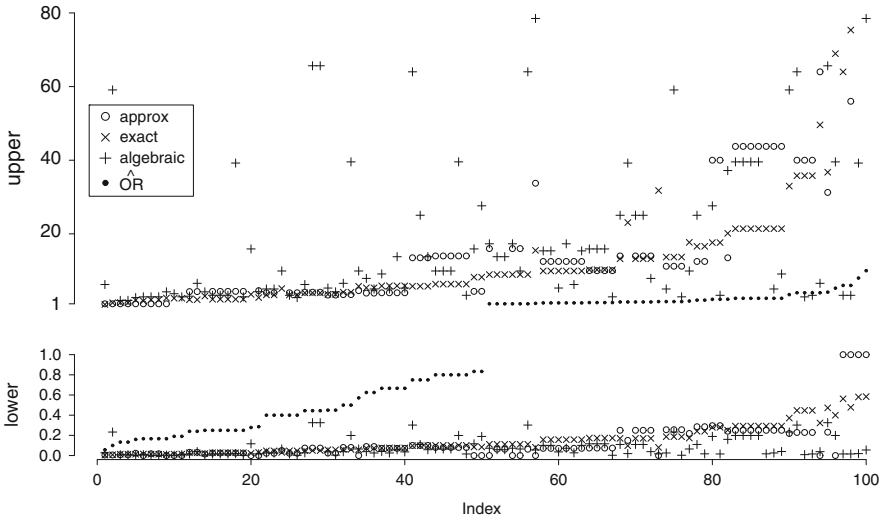


Fig. 1 95%-confidence limits for $n = 15$

4 Simulation Study

In this simulation study we evaluate the performance of the new algebraic confidence interval and compare it to the corresponding exact and asymptotic procedures. The computation is done is `R 2.8.1` and `SAS 9.1`. For the algebraic confidence interval we generate a Markov chain with $l = 500{,}000$ states, disregard the first 50,000 tables in the burn-in-phase and sample each $s = 100$th data set.

We base the simulation on an independence model with probabilities as given in Table 1, hence $OR = 1$. We simulate 100 tables each for a small sample size of $n = 15$, where the approximation of the distribution of $\log(\widehat{OR})$ might not be adequate, and for a relatively large sample size of $n = 50$. Only those generated contingency tables are treated, which lead to strictly positive estimated cell probabilities. For $n = 15$ we had to generate 103 to obtain 100 appropriate data tables, for $n = 50$ no extra tables were necessary. For each simulated table we calculate the approximate, the exact, and the algebraic 95%-confidence interval. For a better representation of our findings, we arrange the results according to the size of the estimated odds ratio. Starting with $n = 15$ in Fig. 1 we observe that the lower confidence limits almost agree for all applied procedures. Upper values of the approximate CI (110.26, 171.20) for the two largest estimated OR's are not shown. The algebraic upper confidence limit reaches a value of infinity in 16 cases. Due to the small sample size too many data tables with cell entries n_{12} or n_{21} equal

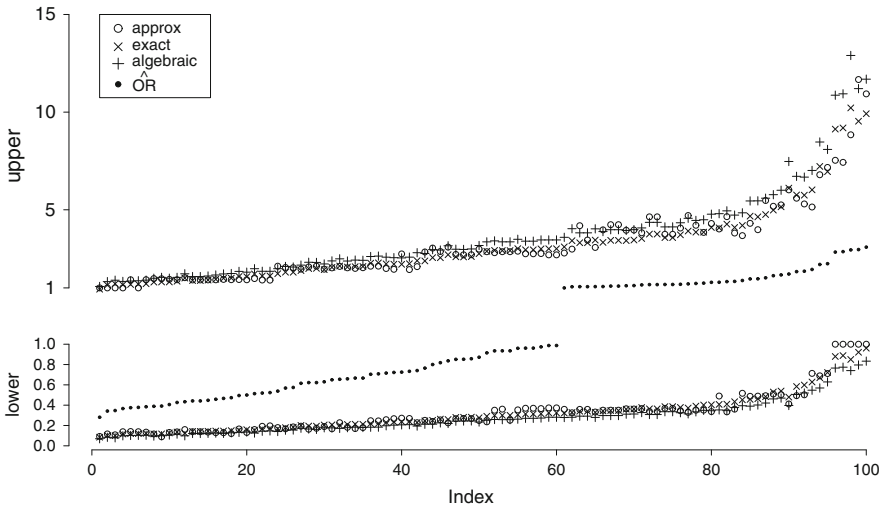


Fig. 2 95%-confidence limits for $n = 50$

to zero are generated in the Metropolis-Hastings algorithm. The modified estimator $\widehat{OR}_{\text{mod}} = \frac{(N_{11}+0.5)\cdot(N_{22}+0.5)}{(N_{12}+0.5)\cdot(N_{21}+0.5)}$ might improve this behavior. Nevertheless is the exact upper limit in 79 cases larger than the approximate or the algebraic counterpart. The exact procedure is conservative, as expected. With the larger sample size of $n = 50$ the lower confidence limits almost coincide (cf. Fig. 2) whereas the upper limits differ. In 48 cases the algebraic confidence interval is the smallest of all three. Altogether we observe that the exact approach usually provides the largest confidence intervals while the algebraic and asymptotic confidence limits are close by.

5 Example

We examine data from a case-control study conducted in the 1980s in Barbone et al. (1993). Its purpose was to analyze the effect of diet on endometrial cancer. Therefore, 103 cases and 236 controls filled out a questionnaire concerning their eating habits. Of interest is the association between consumption of milk and endometrial cancer. The data are given in Table 2.

Since the estimated odds ratio takes a value of 0.663 we assume a preventive effect of milk consumption on endometrial cancer. Inspecting the 95%-confidence intervals we cannot statistically verify an association at level $\alpha = 0.05$ (approximate: 0.411, 1.072; algebraic: 0.389, 1.062; exact: 0.400, 1.106).

Table 2 Observed number of endometrial cancer patients and controls

	Regular consumption of milk	
	Yes	No
Case	61	42
Control	162	74

6 Discussion

In this paper, we show how an alternative algebraic confidence interval for the odds ratio can be derived by using MCMC methods and computational commutative algebra. Traditional and new confidence intervals are compared in a simulation study, where none sticks out as best. However, we expect that the MCMC simulated CI will turn out to be a useful amendment in cases of higher dimensional cases to which the approach can now readily be extended and where exact intervals can not be calculated easily. We further suggest to examine the use of the modified estimator for the odds ratio for 2×2 data tables as well as the Mantel-Haenszel-estimator for the odds ratio for higher dimensional contingency tables in future.

Acknowledgements The financial support of the Deutsche Forschungsgemeinschaft (SFB 475: "Reduction of Complexity for Multivariate Data Structures" and Graduiertenkolleg "Statistical modelling") are gratefully acknowledged.

References

Agresti, A. (2002). *Categorical data analysis*. New York: Wiley.

Barbone, F., Austin, H., & Partridge, E. E. (1993). Diet and endometrial cancer: A case-control study. *American Journal of Epidemiology, 137*, 393–403.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings-Algorithm. *The American Statistician, 49*, 327–335.

Diaconis, P., & Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics, 26*, 363–397.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions*. Hoboken: Wiley.

Haldane, J. B. S. (1955). The estimation and significance of the logarithm of a ratio of frequencies. *Annals of Human Genetics, 20*, 309–311.

Krampe, A., & Kuhnt, S. (2009). Model Selection for Contingency Tables with Algebraic Statistics, in Gibilisco, P., Riccomagno, E., Rogatin, M.-P., Wynn, H.P. (Eds.), *Algebraic and Geometric Methods in Statistics*, Cambridge: Cambridge University Press, 83–97.

Mehta, C. R., Patel, N. R., & Gray, R. (1985). Computing an exact confidence interval for the common odds ratio in several 2×2 contingency tables. *Journal of the American Statistical Association, 80*, 969–973.

Rapallo, F. (2003). Algebraic Markov Bases and MCMC for two-way contingency tables. *Journal of the American Statistical Association, 30*, 385–397.

Riccomagno, E. (2009). A short history of algebraic statistics. *Metrika, 69*, 397–418.

Sørensen, D., & Gianola, D. (2002). *Likelihood, Bayesian, and MCMC methods in qualitative genetics.* New York: Springer.
Woolf, B. (1955). On estimating the relation between blood group and disease. *Annals of Human Genetics, 19*, 251–253.
Zelen, M. (1971). The analysis of several 2×2 contingency tables. *Biometrika, 58*, 129–137.

The CHIC Analysis Software v1.0

Angelos Markos, George Menexes, and Iannis Papadimitriou

Abstract In this paper we describe CHIC (Correspondence & HIerarchical Cluster) Analysis, a specialized software package for Correspondence Analysis-CA (Simple and Multiple) and Hierarchical Cluster Analysis (Benzécri's chi-square distance, Ward's linkage criterion). The implementation of CA is in line with both the French approach and the Gifi System of data analysis. CHIC Analysis combines the graphical interface features of CodeGear Delphi with the computational power of MatLab. The software was implemented as an attempt to contribute to the effectiveness and reliability of CA. For this purpose, it offers a variety of aids to the results' interpretation and tools for the construction of special data tables. A modified version of the CA algorithm is implemented in the multivariate case. Special emphasis has been put on the graphical options for biplots, maps and dendrograms.

1 Introduction

Correspondence Analysis (CA) is a multidimensional data analytic method, suitable for graphically exploring the association between two or more, non-metric variables without a priori hypotheses or assumptions. Similar to Principal Component Analysis, CA results in elegant but simple lower-dimensional displays, so that the principal dimensions (usually two or three) capture the most variance (or inertia) possible. There are two popular ways to treat CA; the geometrical point of view of the French school of data analysis (Benzécri 1992) and the optimal scaling framework of the Gifi System (Gifi 1990).

A common practice among researchers and practitioners is the complementary use of CA and a hierarchical cluster analysis (HCA) procedure, based on Ward's minimum-distance criterion and Benzécri's chi-square distance (Benzécri 1992; Lebart 1994). This specific Ward clustering provides a decomposition of inertia with respect to the nodes of a dendrogram, analogous to the decomposition in the

A. Markos (✉)
Department of Applied Informatics, University of Macedonia, Greece
e-mail: amarkos@uom.gr

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_44,

CA context (Greenacre 2007). More details on the theoretical background of HCA, CA and various extensions can be found in Benzécri (1992), Gifi (1990), Greenacre (1984), Greenacre (2007), Greenacre and Blasius (2006), Murtagh (2005).

CA has become increasingly popular over the last decades and simple and multiple CA were introduced into most of the mainstream statistical software packages. General purpose software such as SAS (SAS Institute Inc. 2003), SPSS (Meulman and Heiser 2005) and XL-STAT (Addinsoft 2007), implement CA offering a variety of options. However, apart from the XL-STAT software, none of the major programs offers recent developments (Nenadic and Greenacre 2006). Additionally, the widespread adoption of **R** (R Development Core Team 2007) within the statistics community led to some important open source CA implementations. The **ca** package (Nenadic and Greenacre 2006) provides functions for Simple, Multiple and Joint CA. Simple and Canonical CA are provided by **anacor** (Mair and de Leeuw 2009), a package which offers alternative plotting options and scaling methods. Multiple CA also known as Homogeneity Analysis (HOMALS) along with various Gifi extensions can be computed by means of the **homals** package (Mair and de Leeuw 2009). **FactoMineR** performs CA (simple and multiple) offering a variety of interpretation options (Le et al. 2008). For most **R** packages a strong level of familiarity with the command line is kind of assumed.

In this paper we present CHIC (Correspondence & HIerarchical Cluster) Analysis, a specialized software which implements CA (Simple and Multiple) and HCA as a complementary method. The software combines two different development tools; Codegear Delphi 7, a visual programming language and MATLAB (The MathWorks, Inc. 2007), a high-level scripting language. This scheme offers a high degree of flexibility since MATLAB is useful for implementing matrix computations, while Delphi offers a variety of tools for the design of graphical user interfaces. The implementation of CA is accompanied by a variety of options for empirical interpretation, statistical inference and visualization, inherent either in the Gifi System or the French approach. Moreover, it offers a modification of the main CA algorithm, so that the analysis of "tall" data sets (objects $\gg$ variables) becomes both feasible and effective. Finally, it is important to note that the development of CHIC Analysis was motivated by the need to teach CA and related methods to students with little or no statistical background and familiarity with the command line.

The paper is organized as follows: Sect. 2 describes in brief the data entry and data management options. The various interpretation options and relative criteria for simple and multiple CA, available in CHIC Analysis, are described in Sects. 3, 4 and 5. A hierarchical clustering procedure as a complementary method is discussed in Sect. 6. A numerical example is given to demonstrate the use of new or interesting features. The paper concludes in Sect. 7.

2 Data Entry and Data Management

CHIC Analysis offers a customized data spreadsheet for direct data entry in the form of either a raw data table (*observations* $\times$ *variables*) or a contingency table of variable categories. Additionally, data can be imported into the spreadsheet from

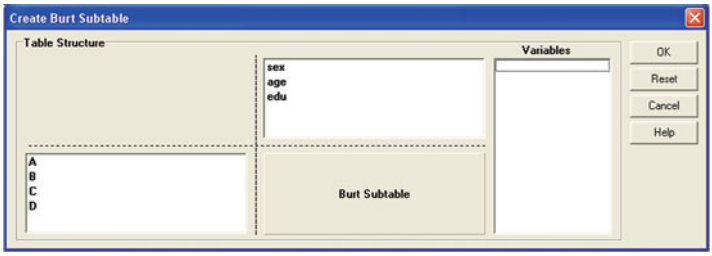


Fig. 1 A tool for Burt Subtable construction

an MS Excel or a text delimited file. There are, also, tools for the discretization of quantitative variables into ordinal and the recoding of categorical variables.

Additionally, the software offers a graphical tool for the direct construction of a Burt subtable or a two-way pivot table, in an abstract form (see Fig. 1). The user can select variables of interest from a list of all available variables in the data set and then drag and drop the desired variables into the available row and column lists. The corresponding variable categories correspond to the rows and columns of a Burt subtable, a contingency table which will be subsequently analyzed by CA.

3 Simple Correspondence Analysis

The implementation of the CA algorithm follows that of Blasius and Greenacre (1994); Greenacre (2007) and its crucial step is the Singular Value Decomposition (SVD) of the standardized residuals matrix. The standard output of CA contains the eigenvalues and the relative and cumulative percentages of explained inertia for all available dimensions. The selection of significant axes can be based on the scree plot and on three different statistical significance tests of the principal inertias proposed by Nishisato (1980), Van de Geer (1993) and Greenacre (1984, 2007), respectively. Table 1 shows the output of CA on the *smoke* dataset, which contains frequencies of smoking habits for staff groups in a fictional company (Nenadic and Greenacre 2006; Greenacre 2007). According to the first two criteria, only the first principal component is statistically significant at the 5% level ($p(VdG) < 0.05$, $p(Nish.) < 0.05$). The options for row and column points include principal coordinates with respect to the dimensionality of the solution, total quality (QLT), inertias (INR), masses (MASS), squared correlations (COR) and contributions (CTR) (see Greenacre 2007 for more details about these concepts). Additional significance criteria of individual points include correlations (SQCOR), which is the equivalent of the factor loadings in PCA (Blasius and Greenacre 1994) and the Best index, which, similar to CTR, is an indicator of which points best explain the inertia of each dimension (SAS Institute Inc. 2003). In the case of supplementary variables, an (S) is appended to the supplementary variable names in the output, which includes only the QLT, INR and COR indices. Table 2 exhibits the row output for

Table 1 Eigenvalues, percentages of inertia and statistical significance

Axis	χ^2	df (VdG)	p (VdG)	χ^2 (Nish.)	-df (Nish.)	p (Nish.)	Inertia	%	Cum. %
1	14.429	6	0.025	14.608	6	0.024	0.075	87.756	87.756
2	1.933	4	0.748	1.893	4	0.755	0.010	11.759	99.515
3	0.080	2	0.961	0.078	2	0.962	0.000	0.485	100.000

Critical χ^2 value $= 15.24$ ($\alpha = 0.05$)

Table 2 Principal row projections, contributions and correlations

	QLT	MASS	INR	Best	F1	INR1	COR1	SQCOR1	CTR1	Best1
SM	0.092	0.057	0.003	3	−0.066	0.000	0.092	−0.304	0.003	0
JM	0.526	0.093	0.012	2	0.259	0.006	0.526	0.726	0.084	0
SE	1.000	0.264	0.038	1	−0.381	0.038	1.000	−1.000	0.512	1
JE	0.942	0.456	0.026	1	0.233	0.025	0.942	0.971	0.331	1
SC	0.865	0.130	0.006	2	−0.201	0.005	0.865	−0.930	0.070	0

the first significant axis of the *smoke* dataset. Optionally, the user can ask for the expected frequencies, three kinds of residuals (plain, standardized and adjusted) of the original contingency table, variable chi-square contributions and the reconstructed input data for a given dimensionality of the solution, as described in Blasius and Greenacre (1994).

4 Multiple Correspondence Analysis

Multiple CA is in fact a simple CA that can be carried out in terms of the SVD on either the indicator matrix or the Burt matrix, a choice which depends on the purpose of the analysis. The indicator matrix is a binary representation of the different categorical values of each variable, while the Burt matrix is equal to the cross-product of the indicator matrix and concatenates all two-way cross-tabulations between pairs of variables (Greenacre 1984, 2007). In cases where a CA on the indicator matrix $\mathbf{Z}$ is of interest, we perform alternatively the SVD on the standardized residuals matrix, calculating on the Burt matrix $\mathbf{B}$. Then, we use the well-known transition formulae of CA (Greenacre 1984, 2007) and the relation between $\mathbf{Z}$ and $\mathbf{B}$, to obtain the results of the CA on $\mathbf{Z}$. This scheme bypasses the decomposition of $\mathbf{Z}$ and can be efficient in the case of "tall" data sets, where the number of objects is much greater than the number of variables. It is important to note that the same CA solution can be also efficiently obtained by means of an Alternating Least Squares algorithm (ALS), which iteratively minimizes a least-squares loss function (Gifi 1990). A thorough description of the modified CA version and its efficiency can be found in Markos et al. (2009).

The standard output of MCA remains the same as in the simple case (see Sect. 3). Additional options include the summary of contributions (CTR) for each variable and the discrimination measures (Gifi 1990; Meulman and Heiser 2005), an important interpretation option inherent in the Gifi System. Furthermore, test values can be calculated for supplementary variables, as a measure of the significance between a variable and an axis (Lebart 2006). The selection of significant axes can be based on a statistical significance test of the principal inertias, proposed by Nishisato (1980) and on Cronbach's alpha, as a measure of reliability of each principal inertia (Greenacre 2007). Finally, in case the analysis is based on the Burt matrix, two inertia adjustment options are offered for solving the low percentage of inertia problem, proposed by Greenacre (2007) and Menexes and Papadimitriou (2003), respectively. Both options are based on the average inertia in Burt's off-diagonal blocks.

5 Visualization Options

The basic plot in CA and MCA is the symmetric map where both rows and columns are plotted in principal coordinates. Depending on the situation, other types of display are appropriate. This can be set with the normalization options for CA and MCA. Following Nenadic and Greenacre (2006), in Table 3 we give a brief overview over the available options and their meanings.

The first three scaling options lead to a biplot, while the last one results in a symmetric map (Greenacre 2007). The intepretation of biplots is enhanced with the option to draw the biplot axes passing through each row (or column) point, as shown in Fig. 2. The dots represent the intersections of the orthogonal projections of points on these biplot axes. In the case of RPN or CPN, the corresponding biplot is likely to be crowded; in that case the interpretation can be based on a table of distances and correlations (Table 4). The distances are in ascending order and indicate a ranking or ordering of the projected points, while Cos2 indicates the square cosine of the angle between a biplot axis and a position vector of a point. For example, for the biplot axis passing through the point "SM", the distance of the projection of the point "Heavy", on the biplot axis, from the point "SM" is 1.791 and the squared correlation between "Heavy" and "SM" is 0.14.

Table 3 Normalization options in CA and MCA maps

Option	Description
RPN - Row Principal	Rows in principal and columns in standard coordinates
CPN - Column Principal	Rows in standard and columns in principal coordinates
SN - Symmetrical	Row and column coordinates are scaled to have variances equal to the singular values
PN - Principal	Rows and columns in principal coordinates (default)

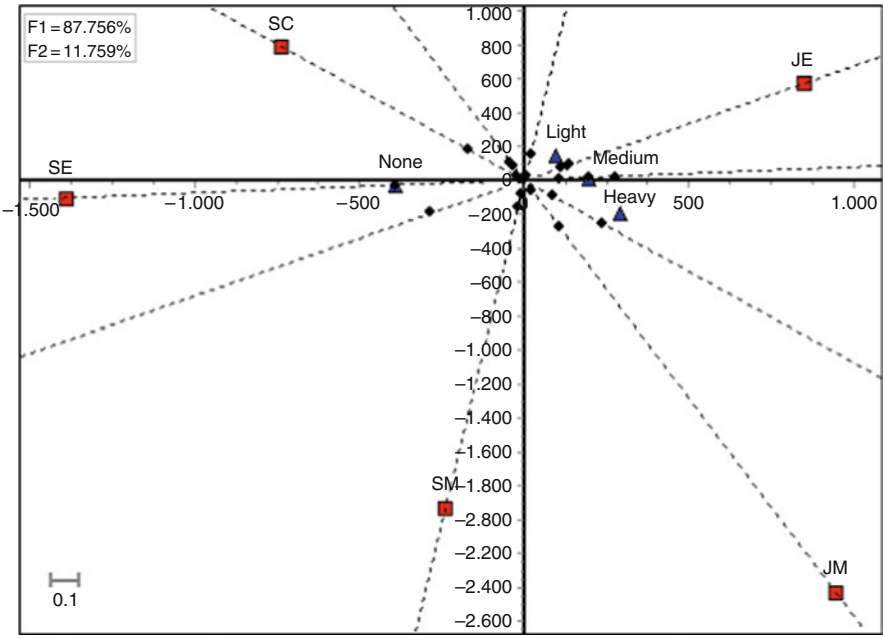


Fig. 2 Asymmetric map of the smoking data with CPN and biplot axes

Table 4 Distances and correlations on biplot axes

SM	Heavy	None	Medium	Light
Distance	1.791	1.873	1.982	2.103
Cos2	0.140	0.039	0.998	0.307
JM	Heavy	None	Medium	Light
Distance	2.469	2.540	2.636	2.744
Cos2	0.390	0.780	0.615	0.109
SE	Heavy	None	Medium	Light
Distance	1.373	1.383	1.403	1.434
Cos2	0.043	0.025	0.754	0.675
JE	Light	Medium	None	Heavy
Distance	0.936	1.009	1.082	1.141
Cos2	1.000	0.874	0.204	0.430
SC	Light	Medium	None	Heavy
Distance	0.987	1.058	1.129	1.187
Cos2	0.404	0.718	0.085	0.949

6 Ward Clustering as a Complementary Method

Hierarchical Cluster Analysis (HCA) can be used as a complementary method to CA, in order to identify relatively homogeneous clusters either in the original data or in the low-dimensional space. A Ward clustering procedure takes into account the chi-square distances between the profiles and the associated masses. This way it provides a decomposition of inertia with respect to the nodes of a dendrogram, analogous to the decomposition of inertia with respect to principal axes in CA (Greenacre 2007). The total inertia (or equivalently the chi-square statistic) of the table is reduced by a minimum at each successive level of merging of the rows (or columns). More details on the HCA algorithm implementation can be found in Greenacre (2007), Lebart (1994), Murtagh (2005).

Furthermore, CHIC Analysis offers a group of interpretation options traditionally called VACOR, which allow the user to explore the representation of clusters derived from the hierarchical trees in factor space and describe the cluster dipoles which take account of the cluster components. More details on the VACOR implementation can be found in Benzécri (1992); Murtagh (2005).

7 Summary

We have presented CHIC Analysis, a specialized software for simple, multiple correspondence analysis and hierarchical clustering. The software contains most of the features of present available software packages as well as various new features that are not available elsewhere. Amongst the main advantages of this program is that it is menu driven, available for free and offers a large variety of complementary options to facilitate data interpretation. The included data entry and data management utilities make it possible to handle directly almost any data table, and this gives the user a great deal of flexibility. The software and its user's manual can be downloaded from http://www.amarkos.gr/en/research/chic.

In future releases, we plan to take advantage of the common mathematical foundation of many multivariate data analysis methods, as a basis for incorporating CA variations and related methods.

References

Addinsoft. (2007). *XLSTAT Ű- Statistical software for MS Excel*. URL http://www.xlstat.com/.

Benzécri, J.-P. (1992). *Correspondence analysis handbook*. New York: Marcel Dekker.

Blasius, J., & Greenacre, M. J. (1994). Computation of correspondence analysis. In M. J. Greenacre and J. Blasius (Eds.), *Correspondence analysis in the social sciences. Recent developments and applications* (pp. 53–75). London: Academic Press.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester: Wiley.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.
Greenacre, M. J. (2007). *Correspondence analysis in practice*. Boca Raton: Chapman & Hall/CRC.
Greenacre, M. J. & Blasius, J. (Ed.). (2006) *Multiple correspondence analysis and related methods*. London: Chapman & Hall.
Le, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software, 25*(1), 1–18.
Lebart, L. (1994). Complementary use of correspondence analysis and cluster analysis. In M. J. Greenacre and J. Blasius (Eds.), *Correspondence analysis in the social sciences. Recent developments and applications* (pp. 162–178). London: Academic Press.
Lebart, L. (2006). Validation techniques in multiple correspondence analysis. In M. J. Greenacre and J. Blasius (Eds.), *Multiple correspondence analysis and related methods* (pp. 179–194). London: Chapman & Hall.
Mair, P., & de Leeuw, J. (2009). Simple and canonical correspondence analysis using the R package anacor. *Journal of Statistical Software, 31*(5), 1–18.
Mair, P., & de Leeuw, J. (2009). Gifi methods for optimal scaling in R: The package homals. *Journal of Statistical Software, 31*(4), 1–21.
Markos, A., Menexes, G., & Papadimitriou, T. (2009). Multiple correspondence analysis for "tall" data sets. *Intelligent Data Analysis*, *13*(6), 873–885.
The MathWorks, Inc. (2007). MATLAB – the language of technical computing, Version 7.5. The MathWorks, Inc., Natick, Massachusetts. URL http://www.mathworks.com/products/matlab/.
Menexes, G. & Papadimitriou, I. (2003). Relations of inertia. In M. J. Greenacre and J. Blasius (Eds.), *Abstracts of international conference on correspondence analysis and related methods*, (p. 56), Barcelona.
Meulman, J. J., & Heiser, W. J. (2005). *SPSS Categories 14.0*. Chicago, IL: SPSS, Inc.
Murtagh, F. (2005). *Correspondence analysis and data coding with Java and R*. London: Chapman & Hall.
Nenadic, O., & Greenacre, M. J. (2006). Correspondence analysis in R, with two- and three dimensional graphics: The ca Package. *Journal of Statistical Software, 20(3)*, 1–13.
Nishisato, S. (1980). *Analysis of categorical data: dual scaling and its applications*. Toronto: University of Toronto Press.
R Development Core Team. (2007). *R: a language and environment for statistical computing. R foundation for statistical computing*. Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.
SAS Institute Inc. (2003). SAS/STAT Software, Version 9.1. Cary, NC. URL http://www.sas.com/, 2003.
Van de Geer, J. P. (1993). *Multivariate analysis of categorical data: applications*. Advanced Quantitative Techniques in the Social Sciences, Vol. 3. Newbury Park: Sage.

Part III
Applications

Clustering the Roman Heaven: Uncovering the Religious Structures in the Roman Province Germania Superior

Tudor Ionescu and Leif Scheuermann

Abstract The antique Roman religion in the province Germania Superior is characterized by the large number of different goddesses and gods which cover and represent the whole Roman civilization (Spickermann 2003). By analyzing them, historians have the possibility to reconstruct everyday life in a mostly illiterate time. At this juncture the multitude of sources makes an overall analysis with traditional methods almost impossible. For this reason automated clustering and classification become more and more interesting for researchers in historical sciences. In this study data consisting of the descriptions of altars (inscriptions, reliefs, and statues), gods and groups of gods, and the towns in the German Southwest where these findings have been made were used in a two-way multivariate cluster analysis. More than 90 gods and over 500 altars from around 100 towns have been initially investigated. Our results consisting of a tree of gods and a tree of towns indicate that there are at least two dominant and distinct religious cults represented by military and imperial gods, respectively, while the tree of towns reveals some interesting structures based on geographical proximity and economic ties.

1 Introduction

In the reconstruction of everyday life of the ancient Romans the main problem lies in the lack of literal sources. Historians have to find another way to gain knowledge about the worries and joys of the ancient Roman. One attempt to reproduce the Roman environment is to reconstruct its religion and belief. The Roman religion, in its diversity, is characterized by an understanding of the Divine as manifesting or even existing in the world. The ancient deity embodies one or more aspects of *the experienced*. The God and the experienced form a unity. For example, Mars embodies aggression and destructiveness. This may reveal itself to believers during

T. Ionescu (✉)
IKE, Universität Stuttgart, Stuttgart, Germany
e-mail: tudor.ionescu@ike.uni-stuttgart.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_45,

the sensual experience of a river or in the middle of a battle. The situations may differ from case to case but the divine does not. The Roman world, i.e. the perceived reality, was not mundane but sacred. The Romans were in each moment surrounded by religion as a vital part of everyday life displayed in the precise attendance of rituals and taboos. As (Rüpke 2006) noted, this also means that the Roman religion revealed to us by places of cult, inscriptions, sculptures, and other archaeological findings can be understood as a mirror of society and everyday life. By analyzing these findings historians have the possibility to get a deeper understanding of everyday life in a mostly illiterate time. The variety and multitude of religious sources always leads to the necessity of a reduction of the complexity of the themes. For traditional analyses this means that the types of gods or the investigation area must be narrowed down and thus the overall scope is lost even before the detailed analysis begins.

In this study we analyzed a total number of 500 inscriptions and sculptures found at 96 different locations (towns) within an area of about 5,000 square kilometers located in today's South-Western Germany. This was a region of mainly geopolitical importance for the Roman province Germania Superior from 90 AD until 360 AD. A total number of 90 names given to 45 distinct types of gods can be found on these artifacts which are sometimes difficult to distinguish because of the similarity in iconography. These data were filtered and prepared for a two-way multivariate cluster analysis. The number of occurrences of the different gods (i.e., mentioned by name or a pictogram) on findings from different locations (towns) have been counted and summarized in a contingency table. The gods were clustered by using the finding locations as variables and vice-versa.

Some of the questions related to these findings which we seek to answer through cluster analysis are the following: Can a schema of relating religious cults be built for the studied area starting from these data? What administrative, social, economical, etc. structures within the area are displayed by the findings? Are the collected archaeological data suitable for cluster analysis and can historians benefit from its results? Our results indicate that there are at least two dominant and distinct religious cults in Germania Superior represented by military and imperial gods, respectively. The key for obtaining a meaningful tree of gods was to use a binary version of the contingency table (i.e., presence–absence data) for clustering the gods. Concerning the tree of towns, we identified two clusters corresponding to two *Civitae* and an agglomeration of towns with economical importance. However, unlike in the case of the tree of gods, these structures could not be validated through correspondence analysis.

2 Data and Methods

The gods mentioned on artifacts together with the towns where they were found constituted the data for our two-way multivariate cluster analysis. Figure 1 shows the related contingency table used as the basis for reconstructing the tree of towns. We

Towns \ Gods	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
StuttgartBadCannstatt	17	0	8	1	4	0	4	1	4	0	10	1	7	5	3	2	0	1	1	4
HeilbronnBoeckingen	3	3	0	1	0	0	0	2	0	1	1	0	1	1	0	1	0	0	1	0
Koengen	10	0	0	0	1	1	0	0	2	0	1	0	3	1	0	1	1	0	0	0
Walheim	9	3	1	0	2	2	0	3	3	0	3	0	4	1	2	3	2	0	2	1
BadWimpfen	4	0	0	4	5	1	0	4	1	0	7	1	0	5	1	0	2	1	0	6
RottenburgamNeckar	12	1	2	0	6	0	0	0	3	0	2	1	2	1	3	3	0	1	0	3
Bietigheim	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0
Boeblingen	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0
Gueglingen	0	0	0	0	1	0	0	0	0	0	2	0	1	0	1	0	0	0	0	1
HausenanderZaber	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0
LoechgauWeissenhof	2	0	0	0	1	0	0	0	0	0	1	0	0	0	2	0	0	0	0	0
Mainhardt	8	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0
Moeglingen	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1
Mundelsheim	2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Neuenhaus	1	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0
NeuenstadtamKocher	0	0	1	0	0	0	0	1	0	0	1	0	0	3	0	2	0	0	0	0
NeuhausenaufdenFildern	1	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0
Oberriexingen	2	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
Oehringen	4	0	0	0	5	1	0	4	2	0	3	3	3	2	2	1	0	0	0	0
Pleidelsheim	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
Pliezhausen	2	0	0	0	0	0	0	0	1	0	3	0	1	0	0	0	1	0	0	0
Schorndorf	1	0	0	0	0	0	0	0	1	0	3	0	0	1	0	2	0	0	0	1
Sindelfingen	0	1	0	0	0	0	0	0	1	0	2	0	2	0	0	0	1	0	0	0
WeilimSchoenbuch	3	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
Welzheim	4	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
StuttgartZazenhausen	1	3	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
SteinheimanderMurr	0	0	0	0	1	0	0	1	1	0	1	0	0	1	1	0	0	0	0	1
Hemmingen	0	0	0	0	1	0	0	0	0	0	2	0	0	0	1	0	0	0	0	1
Gundelsheim	3	0	0	0	1	0	0	0	0	0	2	0	0	1	1	0	0	0	1	2
Murrhardt	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Jagsthausen	7	0	0	0	1	0	1	6	0	0	0	0	1	3	2	0	0	0	2	4
MarbachBenningen	2	3	0	0	4	1	0	1	1	1	2	0	0	4	2	0	2	1	1	1
Waiblingen	2	0	0	0	1	1	0	0	0	0	2	0	2	0	1	0	0	0	1	1

Fig. 1 Contingency table of 33 towns and 20 gods (from left to right): Jupiter, Herecura, Mithras, Matrones, Minerva, Vulkan, Alle Gottheiten, Fortuna, Diana, Campestres, Merkur, Nymphen, Epona, Genius, Herkules, Apollo, Viktoria, Silvanus, Mars, Juno

considered only those towns where more than two artifacts were found. Conversely, only those gods were considered which were mentioned by inscriptions on artifacts from at least two different locations. For computing the dissimilarity matrix for the tree of towns the squared Euclidian distance was used.

For constructing the tree of gods the binary version of the contingency table from Fig. 1 was used. This choice is founded on the argument that the co-occurrence of two or more gods in different towns represents a more appropriate measure of similarity between gods when the actual number of mentions per god for each town is not taken into account. A similar approach for clustering ancient graves based on the presence or absence of certain artifacts in the graves can be found in Manly (1996) while another similar approach for clustering species of fish based on their presence or absence in a selection of lakes can be found in Jackson et al. (1989). The 12 similarity coefficients for ordinal data conveniently listed in Warrens (2008) were then used to compute dissimilarity matrices for constructing the tree of gods.

During the analysis we evaluated the trees reconstructed by 4 different distance-based hierarchical clustering methods: Unweighted Pair Group Method with Arithmetic Mean (UPGMA) (Sneath and Sokal 1973), Neighbor-Joining (NJ) (Saitou and

Nei 1987), Fitch (Felsenstein 1997) from the PHYLIP package (Felsenstein 2005), and Minimum Tree Cost Quartet Puzzling (MTCQP) (Ionescu et al. 2010) from the CLUSTIO package (Ionescu 2008). UPGMA is a fast algorithm which starts by merging the two elements being clustered closest to each other according to a dissimilarity matrix into a new virtual element whereby the distance between the newly emerged element and the remaining unbound elements is recomputed using the arithmetic mean. NJ differs from UPGMA only in the way this distance is recomputed. Fitch starts from a star tree of 3 randomly chosen elements being clustered and attaches the remaining ones, one by one, until there are no unbound elements left. At each step, for each edge of the intermediary tree the sum of squares between the computed (i.e., from the distance matrix) and the realized (i.e., by summing up the edge lenghts for each path of the intermediary tree) distances is evaluated and the element is connected to the edge which yields the smallest sum of squares. MTCQP differs from Fitch in that it uses the least squares criterion on full trees priorly reconstructed using the same additive procedure but a different local cost function based on the 4-point condition applied to quartets of elements being clustered.

Using the 12 coefficients and the 4 clustering methods 48 different trees of gods were reconstructed. The task of objectively evaluating the quality of the clusters formed in each of these trees by visual means proved to be almost impossible. Therefore we applied the majority rule (extended) consensus method (Morris and Powers 2008) to find the best one out of the 12 trees reconstructed by each of the 4 clustering methods.

Finally, using the 4 clustering methods and the squared Euclidian distance matrix 4 trees of towns were reconstructed. The most historically meaningful tree was chosen by means of visual inspection.

3 Results and Interpretations

Figure 2 depicts the consensus tree of gods corresponding to both MTCQP and Fitch and the tree of towns reconstructed by MTCQP. The topology of the tree of gods from Fig. 2 is identical to the ones of the trees reconstructed by MTCQP, Fitch, and NJ using the Jaccard coefficient (Jaccard 1912). By comparison with the displayed consensus tree, in the UPGMA tree of gods Mars, Fortuna, and Genius are placed into the cluster with Jupiter, Merkur, Herkules, Minerva, Juno, and Epona. As will be explained later, historically, this is not correct. In the tree of towns as reconstructed by MTCQP and Fitch two clusters corresponding to the two *Civitae* can be identified.

We now give an interpretation to the trees in Fig. 2 which were considered to be interesting from the historical point of view.

The tree of gods shows two large clusters (i.e. imperial gods and military gods), a smaller one (i.e. abstract gods), and a forth one composed of the gods Vulkan, Viktoria, and Diana. It is remarkable that although the imperial gods are common all over the Roman Empire none of the local gods belongs to this group. This is due

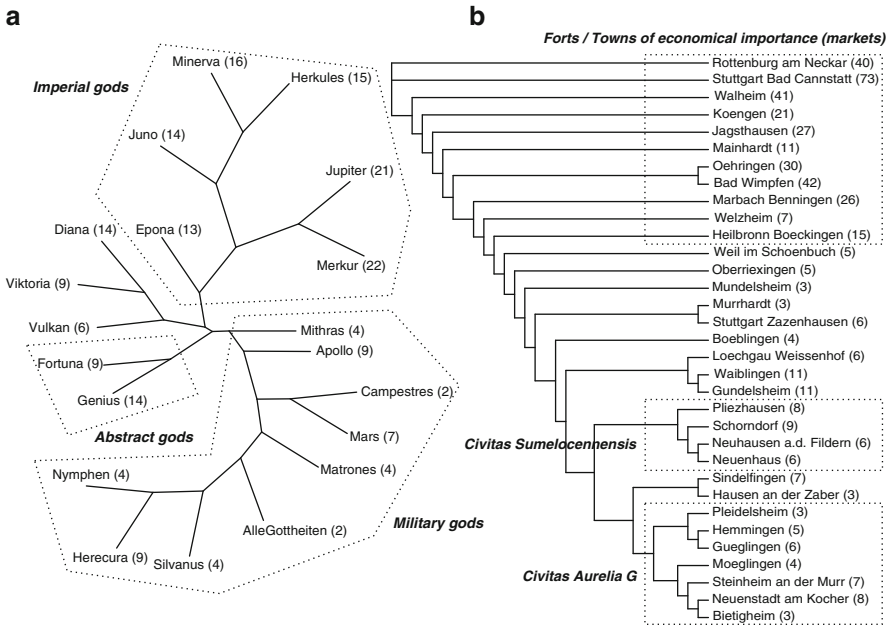


Fig. 2 (**a**) The consensus tree of gods based on the trees reconstructed by MTCQP and Fitch; (**b**) The tree of towns as reconstructed by MTCQP. The values in parentheses represent the number of mentions per god and town, respectively

to the fact that, contrary to the Rhine area, by the time of the Roman occupation there was actually no indigenous population in the studied area so that no local cult could have been passed on to the Romans. This entire cluster of imperial gods is mostly found in a provincial cult called *Jupitergigantensäulen* representing the right order of the state. Herein Jupiter and Mercur take an exceptional position forming a sub-cluster due to their large number of mentions.

The second large cluster is an assemblage of military gods. There is Mars (the god of aggression and war) in close connection to the Campestres – the campus goddesses – as well as the Nymphs which were worshiped on the big aqueducts built by troupes to provide the forts with fresh water. In this cluster there are also mentions of all gods – *Alle Gottheiten* (lat. Ceteri omnes di et deae) which are mostly worshiped by *beneficiarii* – soldiers who were detached from their troupes for civil-administrative and police duties. Herecura is a subtype of Isis / Magna Mater, who had a big temple in Mainz / Mogontiacum – the provincial capital of Germania Superior. The temple was mostly used by the *centuriae* stationed there. In these two cases, the connection to the military forces is evident. A little more difficult to interpret is the appearance of the *matronae* – the mothergodesses worshiped in northern Germania Superior and in Germania Inferior. Many of these findings are sculptures for which a clear assignment to a goddess is almost impossible. Some of them are probably Herecurae or Campestres. This cluster also includes Apollo

and Mithras. In the northern provinces of the empire, Mithras was included into the provincial pantheon (Belayche 2001). Both are related via Sol Invictus – a subtype of Mithras – who represents the sun which is also one of the main aspects of Apollo. The only god not fitting into this cluster is Silvanus.

The last, small cluster consists of only two types of gods: the Genii and the Fortunate. Both are rather collective terms impersonating abstract principles rather than divine entities. Fortuna stands for luck in any way which can be adapted to a specific situation, e.g. Fortuna Redux stands for the lucky return from a journey. Genius is a tutelary god also specialized by the place, time, or principle where the protection shell comes from.

In *the tree of towns* there are two recognizable clusters (*Civitas Aurelia G* and *Civitas Summelocennensis*) and one interesting caterpillar-like tree section containing towns of greater economical importance. These were settlements founded in the direct vicinity to forts and therefore they have strong connections to the military forces. There are many differences among the towns in the length of the period of military occupation and its importance. For example, Rottenburg was probably for only half a year in the 1st century AD a fort whereas settlements like Mainhardt, Öhringen, or Jagsthausen were founded about 150 AD and remained forts until 230/260 AD. Another common feature of these towns is their economical importance which is indirectly connected to the troupes stationed there. With the advance of the military forces, at some of the locations the markets persisted and grew.

The less economically powerful settlements are geographically divided into two clusters. In the first one there are the northern settlements which belong to the administration of the *Civitas Aurelia G* and in the other one there are the southern settlements belonging to the *Civitas Summelocennensis*.

4 Validation of the Results

We consider a validation approach based on the correspondence analysis (CA) method applied to the contingency table from Fig. 1. Figure 3a depicts the CA plot of the gods (with hidden towns) for the binary version of the contingency table from Fig. 1 whereas Fig. 3b shows the CA plot of the towns (with hidden gods) for the exact table from Fig. 1.

The analysis confirms the clustering of the two main groups of imperial and military gods revealed by the tree from Fig. 2a. On the other hand, it reveals that the Genius-Fortuna cluster which appears in the tree is actually a part of the group of military gods and that the Vulkan-Diana-Viktoria cluster is part of the group of imperial gods.

Figure 3b does not allow the clear identification of any clusters. However, the towns forming the *Civitas Sumelocennensis* and the *Civitas Aurelia* which appear as clusters in the tree of towns also appear relatively close together in the CA plot. Stuttgart-Zazenhausen, Murrhardt, and Haus and der Zaber seem to be outliers.

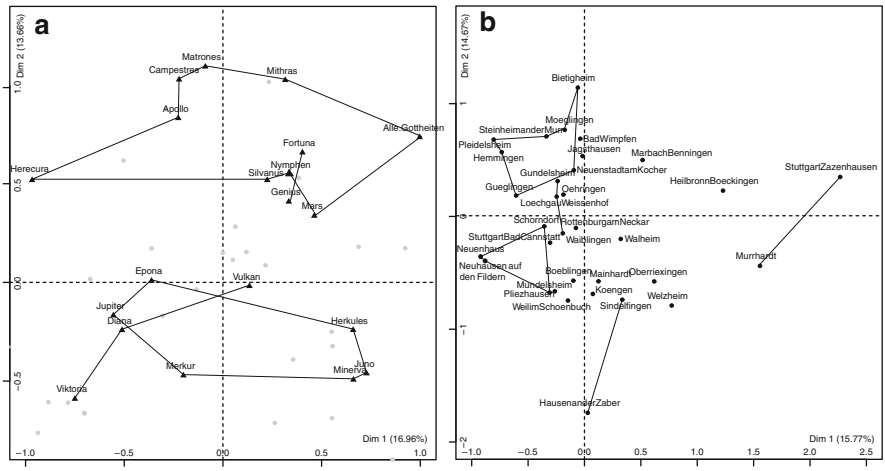


Fig. 3 Results of the correspondence analysis applied for the contingency table from Fig. 1

In conclusion, CA justifies the choice of the binary version of the contingency table for clustering the gods and confirms the two main clusters present in the tree of gods. The fact that there are no clearly distinguishable clusters of towns in the CA plot suggests either that the available data may be insufficient to cluster the towns of Germania Superior using the method presented in this paper or that it is not possible to cluster them solely based on religious artifacts.

5 Conclusion

In this paper we report the results of a two-way multivariate cluster analysis of the data on 45 Roman gods mentioned on archaeological findings from 100 towns in an area of about 5,000 square kilometers in the German southwest. The results reveal at least two dominant and distinct religious cults represented by military and by imperial gods in the tree of gods and two clusters corresponding to two Roman *Civitae* in the tree of towns. The correspondence analysis used to validate the results confirms the structure of the tree of gods while some open questions regarding the tree of towns are left. Was the dataset incomplete or can the localities from an ancient Roman province at all be clustered based on the spreading of religious artifacts used at that time?

We conclude by stating that cluster analysis offers complementary support in interpreting the data on Roman religious cults in the province of Germania Superior. It both provided new insights and confirmed some of the existing hypotheses about the schema of relating religious cults and the administrative, social, and economical structures of that time in the studied area. It should, however, always be used in combination with traditional historical methods of interpretation and with statistical

validation methods. Our further work in this field will be based on a larger database of Roman findings from a territory comprising both Germania Inferior and Superior which will provide more complete data for similar cluster analyses.

References

Belayche, N. (2001). Les cultes syriens dans le Germaines (et les Gaules voisines). In *Spickermann, W. (Hg.): Religion in den germanischen Provinzen Roms* (pp. 285–316). Tübingen: Mohr Siebeck.

Felsenstein, J. (1997). An alternating least squares approach to inferring phylogenies from pairwise distances. *Systematic Biology, 46*, 101–111.

Felsenstein, J. (2005). *PHYLIP (Phylogeny Inference Package) version 3.6*, Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Ionescu, T. B. (2008). *CLUSTIO version 2.1*, Distributed by the author. http://code.google.com/p/clustio.

Ionescu, T. B., Polaillon, G., & Boulanger, F. (2010). Minimum tree cost quartet puzzling. *Journal of Classification*, 27, in press.

Jaccard, P. (1912). The distribution of the Flora in the Alpine Zone. *The New Phytologist, 11*, 37–50.

Jackson, D. A., Somers, K. M., & Harvey, H. H. (1989). Similarity coefficients: Measures of co-occurrence and association or simply measures of occurrence? *The American Naturalist, 133*, 436–453.

Manly, B. F. J. (1996). The statistical analysis of artifacts in graves: Presence and absence data. *Journal of Archaeological Science, 23*, 473–484.

Morris, F. R., & Powers, R. C. (2008). A characterization of majority rule for hierarchies. *Journal of Classification, 25*, 153–158.

Rüpke, J. (2006). In *Die Religion der Römer. Eine Einführung. 2.*, überarb. Aufl. München: Beck.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution, 4*, 406–425.

Sneath, P. H. A., & Sokal, R. R. (1973). In *Numerical Taxonomy* (pp. 230–234). California: W.H. Freeman and Company.

Spickermann, W. (2003). Germania Superior: Religionsgeschichte des römischen Germanien I, In *Religion der Römischen Provinzen 2*, Tübingen, 2003.

Warrens, M. J. (2008). On the indeterminacy of resemblance measures for binary (presence/absence) data. *Journal of Classification, 25*, 125–136.

Geochemical and Statistical Investigation of Roman Stamped Tiles of the *Legio XXI Rapax*

Hans-Georg Bartel, Hans-Joachim Mucha, and Jens Dolata

Abstract Roman stamped tiles of the *legio XXI Rapax* are under investigation coming from different findspots in *Germania Superior*. Their chemical composition was measured by X-ray fluorescence analysis. Here we propose an approach for comparing the measurements coming from different laboratories. First results suggested that the set of tiles can be divided into seven clusters by the *Ward* method (Jain and Dubes 1988). The main part of this paper consists of confirming these results by cluster validation via bootstrapping such as introduced in Dolata et al. (2007). Here a cluster of tiles that contains all findings of Biesheim is of special archaeological interest. The supposed provenance Straßburg-Königshofen can be rejected from the statistical point of view.

1 Introduction

We enlarged our statistical analysis of Roman bricks and tiles by taking into account additional Roman stamped tiles from *Vindonissa* and from other findspots in Switzerland and France, i.e. from the South of the Roman province *Germania Superior*. In this paper, the tiles under consideration were stamped by the *legio XXI Rapax*. Schematic maps of stamp-type distribution of this *legio* in Northern and Western Switzerland are to be found in Giacomini (2001).

First we are going to explain our approaches for comparing the measurements of two different laboratories (see Sect. 3) and for an appropriate data preparation in view of clustering. From the statistical point of view the set of tiles can be divided into seven clusters by the hierarchical *Ward* method (Dolata et al. 2009). The main part of the paper consists of confirming these results by cluster validation via subsampling (simulations) such as introduced in Dolata et al. (2007) and by applying partitional cluster analysis. A cluster of tiles that contains all findings

H.-J. Mucha (✉)
Weierstrass Institute for Applied Analysis and Stochastics (WIAS), D-10117 Berlin, Germany
e-mail: mucha@wias-berlin.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_46,

of Biesheim (France) is of special archaeological interest with the focus on discovering the unknown location of the military brickyard. In this connection, it will be demonstrated again that cluster validation can also be useful.

2 The Roman Stamped Tiles Investigated by Giacomini

Giacomini investigated the petrographic and chemical characterization of tiles from the findspots in Switzerland and Strasbourg (France) with the aim of finding provenances. The results were published in his Thesis n° 1,346 (Université de Fribourg) in 2001, see Giacomini (2001). Concerning the tiles of *legio XXI Rapax*, the data of the chemical compositions of the analyzed samples reported in Annex E were used for our investigations. In 2005, Giacomini published his recent book about the Roman stamped tiles of *Vindonissa* Giacomini (2005). He and M. Maggetti from the Université de Fribourg kindly support our work that will bring together the results obtained from two different laboratories.

3 Preparation of Data Coming from Different Laboratories

The chemical composition of 157 samples were measured using the X-ray fluorescence analysis (XRF) based on the same technical equipment, a PHILIPS PERL'X-2 machine, at two laboratories: 112 measurements at the *Institut de Minéralogie et de Pétrographie, Université de Fribourg* (sample labeling FG standing for the investor Folco Giacomini), and 45 measurements at the *Institut für Chemie und Biochemie – Anorganische Chemie, Freie Universität* Berlin (Gerwulf Schneider and Małgorzata Daszkiewicz, sample labeling W). Concretely, the following findspots and provenances (*) and their cardinality (in brackets) are documented: Alpnach (13), Avenches (5) Biesheim (18), Haut Vully (1), Joressant (1), Kaisten (4), Neuchâtel (1), Frankfurt-Nied (*) (12), *Petinesca* (3), Rheinzabern (*) (1), Rufenach (1), Seeb (30), Strasbourg (7), and *Vindonissa* (60).

For calibration aims, 10 measurements of Fribourg are included in the set of the 45 ones of Berlin that were analyzed repeatedly. concretely, the milled powder of such a tile was divided into two parts, one was analyzed at the laboratory in Berlin and one in Fribourg. By doing so, there is the hope for a good calibration between the two different laboratories.

The following contents will be considered below: the oxides SiO_2, TiO_2, Al_2O_3, Fe_2O_3, MnO, MgO, CaO, Na_2O, and K_2O (in mass percentage), and the trace elements V, Cr, Ni, Zn, Rb, Sr, Y, Nb, and Ba (in ppm). Thus, altogether 18 variables were measured. As mentioned above, to establish comparability 10 samples from Fribourg were made available to be analyzed again in Berlin. On this basis the factors are determined that will be used to adjust the samples from Fribourg. Let denote the calibration samples of Berlin by $\mathbf{X} = (x_{ij})$ and the corresponding ones

Table 1 Slope (1) and calibration factors (2) of each variable for the measurements coming from the laboratory of Fribourg

	Variable (oxid)								
Result	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO	MgO	CaO	Na_2O	K_2O
Slope	1.00	1.01	0.98	1.01	1.01	0.99	0.98	1.05	1.01
Factor f	**1.00**	**0.99**	**1.02**	**0.99**	**0.99**	**1.01**	**1.02**	**0.95**	**0.99**
	Variable (trace element)								
Result	Ba	Cr	Nb	Ni	Rb	Sr	V	Y	Zn
Slope	1.02	0.96	1.00	1.02	1.10	0.96	0.84	1.10	0.93
Factor f	**0.98**	**1.05**	**1.00**	**0.98**	**0.91**	**1.04**	**1.19**	**0.91**	**1.07**

of Fribourg by $\mathbf{Y} = (y_{ij})$. These matrices consist of $I = 10$ rows (samples) and $J = 18$ columns (variables) each, where the element x_{ij} and y_{ij} provide a value for the jth variable describing the ith object of $\mathbf{X}$ and $\mathbf{Y}$, respectively. For each variable $j, j = 1, 2, \ldots, J$, a factor f_j will be obtained such that the line of best fit based on 10 lines is fulfilled:

$$y_{ij} = (\tan\alpha)_j x_{ij} \ (i = 1, 2, \ldots, 10).$$

Under the assumption of errors in both the y- and the x-values, the line of best fit gives the result (Baule 1966):

$$(\tan(2\alpha))_j = \frac{2\sum_i x_{ij} y_{ij}}{\sum_i x_{ij}^2 - \sum_i y_{ij}^2}.$$

From this one gets

$$(\tan\alpha)_j = \frac{1}{(\tan(2\alpha))_j}(\pm\sqrt{(\tan(2\alpha))_j^2 + 1} - 1). \tag{1}$$

At the end the factor f_j of variable j is

$$f_j = (\tan\alpha)_j^{-1} = (\cot\alpha)_j \ . \tag{2}$$

Table 1 presents the final numerical result of calibration based on the 10 double analyzed samples.

Cluster analysis methods based on Euclidean distances are dependent on the scales of the variables. Thus it is necessary to transform the data because the oxides and trace elements come in quite different scales: either in percent or in ppm. Dividing each value of a variable by its arithmetic mean is a quite simple data preprocessing step. It seems to be appropriate because the coefficient of variation (cv) of the original variable becomes the standard deviation of the transformed one. (The cv is defined as the ratio of the standard deviation to the mean.) Thus, the 'native' variability of each variable is preserved. Moreover, the (graphical and numerical)

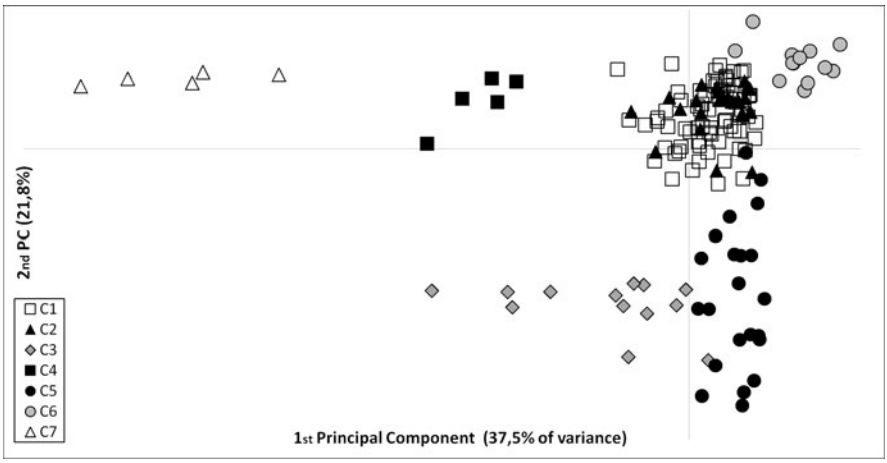


Fig. 1 Principal component analysis plot with cluster membership

comparison of the variables becomes easy because the mean of each transformed variable is equal to 1.

4 Hierarchical Cluster Analysis by *Ward*'s Method

Hierarchical clustering (*Ward*'s sum of squares method, see Jain and Dubes 1988) suggests four, five or seven clusters based on the so-called elbow test (Dolata et al. 2009). From the archaeological point of view the seven clusters solution seems to be particularly suitable for interpretation. Figure 1 shows this solution projected onto the plane of the first two principal components. Also by visual inspection of this plot, seven clusters seem to be a proper partition. In Dolata et al. (2009), this result was described in detail. As seen in Fig. 1, clusters C7 and C4 are isolated. C3 is clearly separated from C5. In the PCA-plot of first and third principal component cluster C6 is more clearly separated, see Dolata et al. (2009). Cluster C5 contains all tiles that come from the findspot Biesheim.

5 Validation of Cluster Analysis Results

As already seen in Dolata et al. (2007), the adjusted *Rand* measure yields reliable results for the investigation of stability of cluster analysis results by successive clustering of bootstrap samples. See, for instance, Hubert and Arabie (1985), Jain and Dubes (1988) and Mucha (2009) for some statistical background of cluster validation based on comparing partitions. First, the simulation algorithm determines the

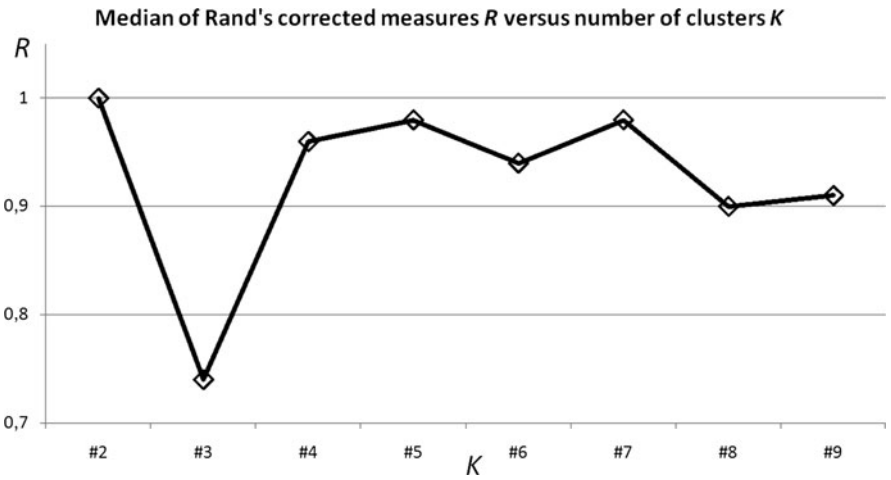


Fig. 2 Adjusted *Rand* measure with respect to the number of clusters

number of clusters, and second it assesses the stability of each single cluster, for details see Hennig (2004) and Mucha (2009). Figure 2 presents the result of validation of the hierarchical *Ward* method by bootstrapping technique using 250 random drawn samples. The maximum adjusted *Rand* measure R (median) is obtained for two clusters. Similar high R-values nearby the maximum are obtained for five, seven and four clusters, respectively. Taking into account the standard deviation S of the R-values for the decision about the number of clusters, the seven clusters solution offers a more than three times lower S than the two class solution. Therefore seven clusters seem to be the more appropriate solution. Similar conclusions hold true for the five and the four cluster solution.

Concerning the validation of each cluster it was found that clusters C5, C6 and C7 are most stable with regard to the following three different statistical measures (see Hennig 2004; Mucha 2009): Jaccard, Dice, and rate of recovery.

Crossing the result of hierarchical cluster analysis (*Ward*'s method, ordinate) and the result of partitional clustering (K-means) gives the contingency Table 2. The latter verifies the former to a relatively high degree (apart from cluster C1). In Mucha et al. (2009), Mucha et al. show that the clusters of the *legio XXI Rapax* remain stable when the data under investigation is extended by tiles stamped by other military units.

Finally, the focus will be on the investigation of the interesting cluster C5 containing all tiles from Biesheim. The archaeologist M. Reddé supposed that Straßburg-Königshofen is a potential provenance for these tiles. To verify this, the total set of tiles from C5 and Straßburg-Königshofen is investigated by *Ward*'s cluster analysis method. Figure 3 shows the principal component analysis (PCA) plot that reflects a clear separation of tiles of C5 from tiles of the identified military brickyard Straßburg-Königshofen.

Table 2 Crossing the results of clustering by *Ward* and *K*-means

		K-means cluster analysis							
	Cluster	K1	K2	K3	K4	K5	K6	K7	Total
Ward's method	C1	**36**		1			**34**		71
	C2				**28**				28
	C3			**13**					13
	C4							**5**	5
	C5		**21**		1				22
	C6					**13**			13
	C7							**5**	5
	Total	36	21	14	29	13	34	10	157

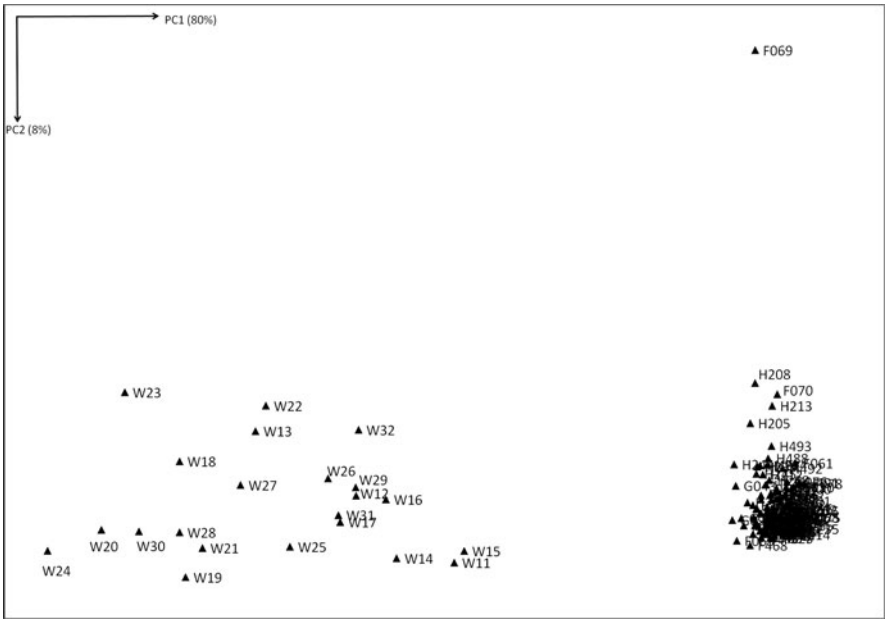


Fig. 3 PCA plot of tiles of cluster C5 (label W) and of tiles manufactured in Straßburg-Königshofen (F, G, and H)

Figure 4 presents the result of validation of the hierarchical *Ward* method by bootstrapping technique. It shows both a graphical representation of the most important statistical results of simulations concerning the adjusted *Rand* index R and a table containing the corresponding numerical values of these univariate statistics. The reading of Fig. 4 is as follows: The axis at the left hand side and the bars in the graphic are assigned to the standard deviation S of R, whereas the axis at the right hand side and the box-plots are assigned to other statistics of R (average, upper and lower 5% quantile). The average of R for $K = 2$ takes the maximum value. That means, the two cluster solution can be confirmed in a high degree for almost all

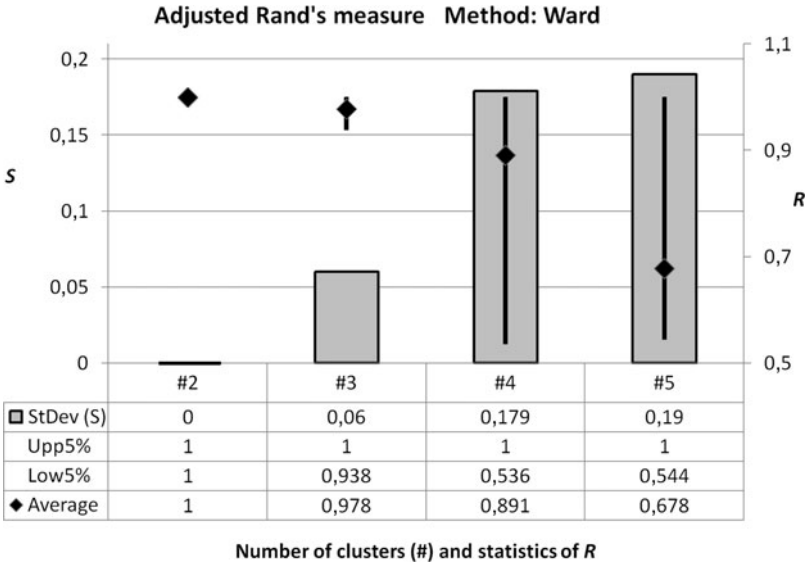


Fig. 4 Adjusted *Rand* measure with respect to the number of clusters

samples. Moreover, the two cluster solution gives the minimum standard deviation equals 0. The higher the number of clusters the lower the average of the adjusted *Rand* index *R* median and the higher the standard deviation become, respectively. Therefore, without any doubt the tiles of cluster C5 were not manufactured in Straßburg-Königshofen.

6 Interpretation of the Geochemical Clusters

From the statistical point of view the set of tiles can be divided into seven clusters obtained by hierarchical cluster analysis. This can be confirmed by validation via bootstrapping such as described in Dolata et al. (2007). Moreover partitional clustering gives similar results. The location of the corresponding military brickyard of the cluster that contains all findings from Biesheim is not yet known and therefore remains an open question. As far as we today know, these tiles were not manufactured in Straßburg-Königshofen. As a result of validations, archaeologists can now modify and consolidate their ideas about the internal structure of Roman tiles stamped by the *legio XXI Rapax*.

References

Baule, B. (1966) *Die Mathematik des Naturforschers und Ingenieurs – Band II: Ausgleichs- und Näherungsrechnung*. Leipzig: Hirzel.

Dolata, J., Bartel, H.-G., & Mucha, H.-J. (2009) Geochemische und statistische Erkundung der Herstellungsorte von Ziegeln der *legio XXI Rapax*. In M. Reddé (Ed.), *Oedenburg – Fouilles françaises, allemandes et suisses à Biesheim et Kunheim, Haut-Rhin, France. Vol. 1: Les camps militaires Julio-Claudiens* (pp. 355–364). Römisch-Germanisches Zentralmuseum, Mainz.

Dolata, J., Mucha, H.-J., & Bartel, H.-G. (2007). Uncovering the internal structure of the roman brick and tile making in Frankfurt-Nied by Cluster Validation. In R. Decker and H.-J. Lenz (Eds.), *Advances in data analysis* (pp. 663–670). Berlin: Springer.

Giacomini, F. (2001). *The roman stamped tiles of Vindonissa (Northern Switzerland): Provenance and technology of the production*. Thesis n° 1346. Université de Fribourg (Suisse).

Giacomini, F. (2005). The Roman Stamped Tiles of Vindonissa (1st Century AD., Northern Switzerland), Provenance and technology of production – an archaeometric study. BAR International Series 1449, Oxford.

Hennig, C. (2004). A general robustness and stability theory for cluster analysis. Preprint no 7, Universität Hamburg.

Hubert, L. J., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*, 193–218.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. New Jersey: Prentice Hall.

Mucha H.-J. (2009). Cluscorr98 for Excel 2007: Clustering, multivariate visualization, and validation. In H.-J. Mucha and G. Ritter (Eds.), *Classification and clustering: models, software and applications* (pp. 14–41). Report no. 26. Berlin: WIAS.

Mucha, H.-J., Bartel, H.-G., & Dolata, J. (2009). Zur Klassifikation römischer Ziegel von Fundorten im südlichen Obergermanien. In A. Hauptmann and H. Stege (Eds.), *Archäometrie und Denkmalpflege* (pp. 144–146). (*Metalla* (Bochum), Sonderheft 2).

Land Cover Classification by Multisource Remote Sensing: Comparing Classifiers for Spatial Data

Alexander Brenning

Abstract Land cover classification is a standard remote-sensing task in which typically multispectral satellite data is used to identify features such as land use. The detection of rock glaciers is a particularly challenging task that requires the combination of satellite data with terrain analysis data because their spectral signature alone is not distinct enough for their classification based on satellite imagery alone. The performance improvements that can be achieved by selecting an optimal classifier in this particular land cover classification problem are investigated.

In the case study, eleven statistical and machine-learning techniques are compared in a benchmarking exercise, including logistic regression, generalized additive models (GAM), linear discriminant techniques, the support vector machine, and bootstrap-aggregated tree-based classifiers such as random forests. Penalized linear discriminant analysis (PLDA) achieves a median false-positive rate (mFPR, estimated by cross-validation) of 8.2% in early detection of rock glaciers at a sensitivity of 70%, which is significantly better than all other classifiers. The GAM and linear discriminant analysis are second best (mFPR: 8.8%). The mFPR of the worst three classifiers is about one-quarter higher compared to the best three classifiers.

The land cover classification problem is further analyzed in general terms from a methodological perspective, highlighting potentials and pitfalls related to phenomena including error estimation in the presence of spatial dependence, high-dimensional problems in hyperspectral remote sensing, and indirect models.

1 Introduction

Land cover classification and change detection are important applications of (mostly supervised) classification techniques in remote sensing, i.e. the study of the Earth's surface using aerial and satellite imagery. As an example, according to Thomson

A. Brenning
Department of Geography and Environmental Management, University of Waterloo, 200 University Ave. W., Waterloo, Ontario N2L 3G1, Canada
e-mail: brenning@uwaterloo.ca

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_47,

Reuters' ISI Web of Knowledge, the staggering number of 577 scientific papers with either 'land cover classification', or 'change detection' and 'remote sensing' in the title, abstract or keywords was published during the 5-year period from 2004 to 2008. However, only 19 of these papers fall within the subject fields of computer science, mathematics or statistics, indicating an imbalance between theoretical and applied research.

Based on these observations, this contribution aims at making the land cover classification problem more well-known among more theoretical researchers. For this purpose, the benefits of using modern classification techniques are demonstrated in the challenging land cover classification problem of rock glacier detection (Sect. 2), and based on this case study and related studies some of the methodological challenges of land cover classification and geospatial analysis are illuminated (Sect. 3).

In land cover classification, the land cover type (e.g., land use, vegetation or crop type, building type) is known in some areas or at individual locations from so-called 'ground-truthing', i.e. based on ground inspection, ancillary data sources or expert interpretation of higher-resolution imagery. The land cover type may be binary, but it is more often polychotomous, sometimes with 10–20 classes. In multispectral remote sensing, predictor variables represent the reflective properties of ground surface at different spectral wavelengths or 'bands'. As an example, Landsat ETM+ provides nine bands of different visible and near-infrared wavelengths at 30 m spatial resolution, but a wide range of systems with different spectral resolutions ($<$10 to $>$100 bands) and spatial resolutions ($<$1 m to $>$1 km) is available. A classifier, often quadratic discriminant analysis (QDA; referred to as maximum-likelihood classification in remote sensing), is then fitted on the training samples – more or less sparse point or area samples – and applied pixel by pixel to a remote-sensing image to predict land cover across a study region.

Much of the remote-sensing literature further focuses on the derivation of enhanced information such as band ratios or texture filters from spectral data to improve the predictive performance (compare Lu and Weng 2007). The comparison of new satellite sensors to older ones as well as the combination of different data source (multisource classification) is also receiving significant interest, although a formal benchmarking framework (Hothorn et al. 2005) is rarely adopted (Brenning 2009). In many applications, terrain analysis variables provide valuable information that helps improve classifications based on remotely-sensed data (Sect. 2 and Brenning 2009; Michelson et al. 2000).

Segmentation methods and object classification are increasingly being used to exploit the spatial characteristics of image data (Lu and Weng 2007). Apart from these geometric classification approaches, remote-sensing research also focuses on enhancing the feature space of the problem at hand, and using advanced statistical and machine-learning techniques (Brenning 2009; Brenning et al. 2006; Chan and Paelinckx 2009; Huang et al. 2002). This paper examines this issue from a methodological perspective, highlighting known and novel links between classification science and remote-sensing applications.

2 Benchmarking Classifiers for Multisource Rock Glacier Detection

The case study focuses on the application of statistical benchmarking approaches (Hothorn et al. 2005) to (1) rank statistical and machine-learning classification techniques and (2) assess the benefit of integrating multispectral remote-sensing data with terrain attributes in the detection of rock glaciers across a mountain area. Rock glaciers are a landform resulting from the creep of ice-rich (∼40–60% by volume) mountain permafrost. They are important stores of frozen water in some dry mountain areas (Brenning et al. 2007), however relatively little is known about their exact distribution and location. Rock glacier ice is not exposed at the surface, making their detection particularly difficult. The use of terrain attributes derived from digital elevation models (DEMs) is a promising approach, and their combination with multispectral remote-sensing data is the objective of the present case study, which is presented elsewhere in detail (Brenning 2009).

2.1 Materials and Methods

The learning sample consists of $N = 2071$ points randomly distributed across the San Juan Mountains, Colorado, USA (2784 km^2). At these locations, rock glacier presence ($n = 86$) and absence was determined visually from high-resolution (1 m) grey-scale air photos. Eleven spectral and derived remote-sensing variables (Landsat ETM+, 30 m resolution) and 54 terrain attributes (Shuttle Radar Topography Mission SRTM, 30 m, and smoothed topographies) are used as predictor variables (Brenning 2009; Brenning et al. 2007). Many of the terrain attributes are highly correlated, and it is known from previous studies that some of their relationships to the response variable are nonlinear; interactive effects have not been examined in this area but can also be expected to exist (Brenning et al. 2007).

The predictive performance of classification techniques is determined in terms of the median false-positive rate (FPR) at a 70% sensitivity using 100-repeated five-fold cross-validation stratified based on the response variable. Differences in FPR were first examined with a global permutation test (Hothorn et al. 2005), and then in pairwise signed-rank tests using a Simes procedure to control the false-discovery rate of 5%.

This study compares 11 different classification techniques ranging from discriminant analysis (DA) methods to generalized (logistic) linear and additive models (GLM, GAM), tree-based ensemble techniques and the support vector machine (SVM). The DA methods studied are linear DA (LDA), stabilized LDA (SLDA) (Läuter 1992) and penalized LDA (PLDA; with default parameter $\lambda = 1$) (Hastie et al. 1995). Logistic regression and the logistic GAM use stepwise forward variable selection based on the AIC; the GAM also uses this criterion to choose between linear and nonlinear (2 d.f.) effects. The Polyclass method is similar to the GAM

and used with default settings (Kooperberg et al. 1997). The tree-based techniques applied in this study are bagging, bundling with SLDA as an ancillary classifier (Hothorn and Lausen 2005), and random forests. Finally, the SVM (C-classification with radial basis function kernels and shrinking heuristics, LIBSVM implementation) is used in two settings: With default hyperparameters ($C = 1, \gamma = p^{-1}$, where $p =$ # of variables), and with an grid-search hyperparameter estimation based on an internal cross-validation (within each cross-validation design set).

The predictive performance of the best three classifiers is further estimated using only terrain attributes (TA) and using only remote-sensing data (RS) in order to assess the benefits of combining both sets of variables.

2.2 Results

The comparison of the 11 classification techniques (Fig. 1) indicates that, in general terms, the less flexible linear and moderately nonlinear methods outperform the more flexible tree-based ones and the SVM. PLDA performs significantly better than all other classifiers, and the GAM and LDA are on rank two but perform significantly better than the lower-ranked methods. The median FPR of the worst-performing three classifiers is 27% higher than the FPR of the best three techniques, i.e. the area incorrectly classified as rock glacier can be considerably reduced by selecting the optimal method.

Interestingly, the SVM with internal hyperparameter tuning (SVM-CV) does not perform better than with fixed default parameters. It appears that the data set is too small to produce stable hyperparameter estimates. PLDA and SLDA attempt to solve the problem of (near-)high-dimensionality in different ways. The dimension reduction approach used by SLDA, which is based on singular value decomposition,

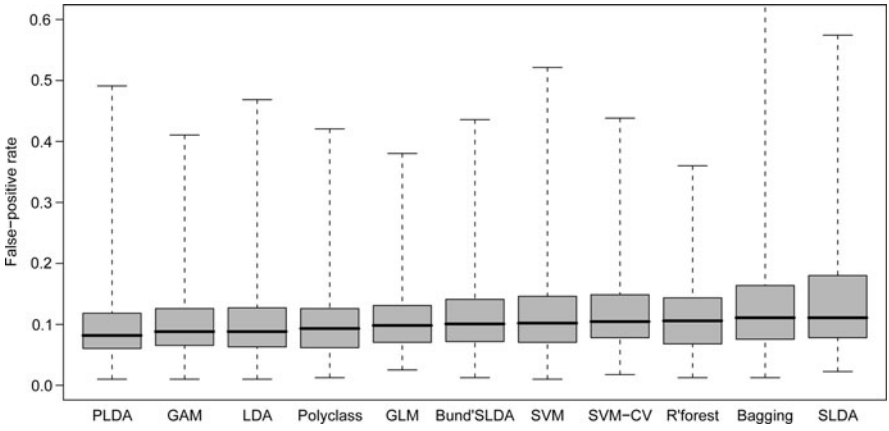


Fig. 1 Cross-validated false-positive rates achieved by the 11 classification techniques at a 70% sensitivity using terrain attributes and remote-sensing data

Table 1 Median false-positive rates achieved by the best three classifiers in rock glacier detection at a sensitivity of 70% using different sets of predictor variables (TA: terrain analysis; RS: multispectral remote-sensing)

Method	TA + RS	TA	RS
PLDA	0.082	0.134	0.222
GAM	0.088	0.145	0.194
LDA	0.088	0.131	0.227

does not effectively retain all information relevant for discriminating the feature of interest. It should also be noted that QDA, one of the standard methods frequently used by remote-sensing practitioners, failed because of the large number of predictor variables compared to the number of presence data.

The multisource classification clearly outperformed classification based on only terrain attributes or only remote-sensing data (Table 1).

3 Discussion

3.1 State-of-the-Art Classifiers for Land Cover Mapping

The case study presented in this paper focused on the use of relatively new classification techniques such as PLDA, GAM, or random forests compared to more traditional ones (LDA, GLM), and on the comparison of predictive performances of different data sources and their combination (multisource classification). In this example, geomorphological expert knowledge tells us that nonlinear, non-monotonous relationships must be present in several predictor variables (e.g. catchment slope), suggesting the use of nonlinear modeling techniques such as the GAM or tree-based techniques. However, less flexible, even linear methods (PLDA, LDA) were better able to predict the response variable, presumably because the availability of a large set of strongly correlated predictors allowed to capture nonlinear effects sufficiently well.

The present case study and similar comparative studies in land cover classification (Brenning et al. 2006) and landslide susceptibility modeling (Brenning 2005) indicate that linear methods may perform well in situations where large numbers of strongly correlated predictor variables are present. Tree-based methods tend to overfit to the training data, which may in some cases be related to spatial autocorrelation (Brenning 2005; Brenning et al. 2006). It is however difficult to generalize these results.

3.2 High-Dimensional Problems in Remote Sensing

High-dimensional or nearly high-dimensional problems occur in a variety of situations in land cover classification, especially (1) when hyperspectral remote-sensing data is used (Chan and Paelinckx 2009), (2) when multitemporal imagery is combined (Brenning et al. 2006), or when multiple data sources such as terrain attributes and remote-sensing imagery are 'stacked' (Brenning 2009). Standard methods designed for 'low-dimensional' problems may not be applicable in these situations, or they may overfit to strongly correlated and partly non-informative variables. A variety of techniques such as PLDA and random forests have been developed for this situation. High-dimensional problems similar to the analysis of hyperspectral remote-sensing data occurred earlier in chemometrics in the context of spectroscopy. Ad hoc dimension reduction approaches such as the use of the most important principal components or eigenvectors do not appear make effective use of the available data, as can be seen in the poor performance of SLDA in the present case study compared to PLDA (Brenning 2009). Dimension reduction by principal component analysis is however still common in applied remote sensing.

3.3 Spatial Error Estimation

A common challenge in modeling spatial data is the autocorrelation of data observed at nearby locations. This phenomenon, which is the starting point for geostatistical theory, may not only render ordinary regression estimators inefficient. In a predictive situation, it may also be necessary to take spatial dependence into account in error estimation using resampling-based methods (Davison et al. 2003). When the learning sample has a high spatial density (which is not the case in the present case study), the standard (non-spatial) cross-validation or bootstrap error rates will be nearly identical to the apparent error because adjacent samples are virtually identical and may end up in the test and training set.

Block bootstrap approaches for dependent data have mainly been studied in the similar situation of time series analysis, but the size and properties of the blocks to be resampled may influence the results (Bühlmann 2002; Zhu and Morgan 2004). A different approach is a spatial cross-validation in which training and test samples are required to have some minimum distance (Brenning 2005). In the context of landslide susceptibility modeling based on densely gridded inventory data, the apparent error rate was heavily overoptimistic compared to the spatial estimator (Brenning 2005).

Special resampling approaches have also been proposed for grouped data. In crop detection, where pixel-level multispectral data is grouped at the field level due to field-specific random effects (e.g. same cultivar and sowing date), it may be advisable to resample the learning sample at the field level rather than the pixel level (Brenning et al. 2006). Similar procedures have been proposed for other grouping structures such as the classification of paired organs (Brenning and Lausen 2008).

It is evident that further research on error estimation in the spatial domain is needed because this has immediate consequences for benchmarking classifiers and data sources such as new satellite sensors.

3.4 Indirect Classification in Remote Sensing

The theoretical framework of indirect classification may provide further new insights into the classification of remote-sensing data and may potentially improve predictive performances (Peters et al. 2005). Indirect classification attempts to make use of ancillary information that is available on the training set but not normally in a predictive situation. This ancillary information is represented by intermediate variables that are predicted based on the available predictor variables and then incorporated into the set of variables of the final model.

In rock glacier detection, the 'non-rock glacier' class corresponds to a number of land cover classes that are not of interest for the final results, but are known on the training set. This provides us with a structured class definition that may be exploited to improve the predictive performance because different predictor variables are useful to predict vegetation and vegetation-free land, both of which are non-rock glacier areas. On the other hand, the use of ancillary models predicting for example ground surface temperature (as a proxy for the presence of permafrost) as an input to permafrost presence/absence models fits perfectly within the framework of indirect models.

At a simpler level, practitioners often use 'thresholding' and 'masking' steps as preprocessing steps. Systematically integrating these steps into indirect land cover classification models would help incorporate the uncertainties of these steps into error estimation, and identify optimal decision thresholds.

4 Conclusions

The present work discussed a number of potentials and pitfalls in the application of classification techniques to land cover mapping. Adequate spatial error estimation and benchmarking methods are needed in order to provide guidance in the selection of classification techniques and feature extraction algorithms, the evaluation of new satellite sensors or the integration of multiple data sources. Most studies however have relied on apparent or test set error rates instead of (spatial) resampling-based methods, and hypothesis tests on differences between classification approaches are therefore not available in most existing comparisons (compare however Brenning 2005; Brenning et al. 2006; Brenning 2009).

Strengthening the interaction between researchers in the field of classification methodology and remote-sensing applications could therefore be very fruitful for both communities. The ifcs 2009 meeting 'Classification as a Tool for Research'

in Dresden with its Special Interest Sessions on Spatial Classification and Spatial Planning provided a platform for this dialogue, which may be extended to other fields of geospatial analysis such as hazard susceptibility modeling (Brenning 2005), digital soil mapping (McBratney et al. 2003), or species habitat modeling (Prasad et al. 2006). Moreover, recent developments in the integration of geographical information systems (GIS) and the statistical software R also promise to boost interdisciplinary cooperation in geospatial analysis. The existing evidence suggests that a reduction of error rates on the order of one-quarter is at stake (Brenning 2005, 2009; Brenning et al. 2006). Remote-sensers would be ill-advised to disregard this potential improvement, which comes at virtually no cost compared to new satellite sensors.

References

Brenning, A. (2005). Spatial prediction models for landslide hazards: Review, comparison and evaluation. *Natural Hazards and Earth System Sciences, 5*, 853–862.

Brenning, A. (2009). Benchmarking classifiers to optimally integrate terrain analysis and multispectral remote sensing in automatic rock glacier detection. *Remote Sensing of Environment, 113*, 239–247.

Brenning, A., Grasser, M., & Friend, D. A. (2007). Statistical estimation and generalized additive modeling of rock glacier distribution in the San Juan Mountains, Colorado, USA. *Journal of Geophysical Research, 112*, F02S15.

Brenning, A., Kaden, K., & Itzerott, S. (2006). Comparing classifiers for crop identification based on multitemporal Landsat TM/ETM data. In *Proceedings, Second Workshop of the EARSeL Special Interest Group on Remote Sensing of Land Use and Land Cover, 28–30 September 2006* (pp. 64–71). Germany: Bonn.

Brenning, A., & Lausen, B. (2008). Estimating error rates in the classification of paired organs. *Statistics in Medicine, 27*, 4515–4531.

Bühlmann, P. (2002). Bootstraps for time series. *Statistical Science, 17*, 52–72.

Chan, J. C.-W., & Paelinckx, D. (2009). Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment, 112*, 2999–3011.

Davison, A. C., Hinkley, D. V., & Young, G. A. (2003). Recent developments in bootstrap methodology. *Statistical Science, 18*, 141–157.

Hastie, T. J., Buja, A., & Tibshirani, R. (1995). Penalized discriminant analysis. *Annals of Statistics, 23*, 73–102.

Hothorn, T., & Lausen, B. (2005). Bundling classifiers by bagging trees. *Computational Statistics & Data Analysis, 49*, 1068–1078.

Hothorn, T., Leisch, F., Zeileis, A., & Hornik, K. (2005). The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics, 14*, 675–699.

Huang, C., Davis, L. S., & Townsend, J. R. G. (2002). An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing, 23*, 725–749.

Kooperberg, C., Bose, S., & Stone, C. J. (1997). Polychotomous regression. *Journal of the American Statistical Association, 92*, 117–127.

Läuter, J. (1992). *Stabile multivariate verfahren*. Berlin: Akademie Verlag.

Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing, 28*, 823–870.

McBratney, A. B., Mendonça Santos, M. L., & Minasny, B. (2003). On digital soil mapping. *Geoderma, 117*, 3–52.

Michelson, D. B., Liljeberg, B. M., & Pilesjö, P. (2000). Comparison of algorithms for classifying Swedish landcover using Landsat TM and ERS-1 SAR data. *Remote Sensing of Environment, 71*, 1–15.

Peters, A., Hothorn, T., & Lausen, B. (2005). Generalised indirect classifiers. *Computational Statistics & Data Analysis, 49*, 849–861.

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems, 9*, 181–199.

Zhu, J., & Morgan, G. D. (2004). A nonparametric procedure for analyzing repeated measures of spatially correlated data. *Environmental and ecological statistics, 11*, 431–443.

Are there Cluster of Communities with the Same Dynamic Behaviour?

Martin Behnisch and Alfred Ultsch

Abstract Demographic and economic changes lead to the phenomena of growing and shrinking cities. The issue of this article is to find groups (cluster) of communities with the same dynamic characteristics in Germany. Community Data Mining represents a methodological approach that discovers logical, mathematical and partly complex descriptions of urban patterns and regularities inside statistical data. The approach relies on 12,430 communities and refers to data from well known and easily accessible institutions. Emergent SOM is presented as an appropriate and powerful method for clustering and classification. The application of U*-Matrix shows that it is of high value, first, to visualize the structure of highdimensional data and second, to detect meaningful classes. Knowledge Discovery is applied to find a description and recognition of a given set of cluster. The structure and the machine generated explanations were validated mindful of the spatial analyst and yielded a spatial abstraction. Such approaches might lead to a benchmark system for regional policy or to other strategic instruments such as semi or fully automated urban monitoring systems.

1 Introduction

The German financial system leads to extensive disparities between different types of communities. It can be assumed that economical crisis and demographical changes will foster regional and spatial disparities. New types of communities are arising and precise individual concepts are needed for their urban development. As communities are facing such complex problems, it is necessary to recognize this complexity and tackle it with comprehensive and multidimensional approaches. There is a need for discovering the properties on which politicians and planners

M. Behnisch (✉)
Institute of Historic Building Research and Conservation, ETH Hoenggerberg, HIT H 21.3, CH-8093 Zurich, Switzerland
e-mail: Behnisch@arch.ethz.ch

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_48,

shall work with. Needless to say that many former studies already discovered growing and shrinking processes in Germany (Gatzweiler et al. 2003; Siedentop et al. 2003). It is therefore well-known that the southern and western part of Germany is growing and the eastern part (former GDR) is shrinking. But these studies were just interested in summarizing growing or shrinking properties, but not in a complex understanding and extraction of multidimensional properties. For this reason the authors want to identify new patterns as well as multiple dynamic behaviours of all communities. The assumption is that unexpected patterns will emerge both in dynamic and in localisation.

2 Data for German Community Dynamics

Six variables were selected for the classification analysis: population, migration, taxing-capacity, dwellings, employment rate and commuter ratio. The standardized and comparable data was created by the Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR). These variables are often used in former approaches (Gatzweiler et al. 2003). A multidimensional dynamic means a combined view on all these variables. The dynamic processes are mostly characterized by positive or negative percentage quotations between the year 1994 and 2004. Some variables are used to describe the present situation referring to one specific year. For example taxing capacity provides an indication of the economical and financial situations of communities (Bundesamt 2005). Table 1 shows the calculation of variables.

The inspection of data includes the visualisation in form of histograms, Q-Q-Plots, PDE-Plots (Ultsch 2003) and Box-Plots. The authors decided to use transformation measurements such as ladder of power to take into account restrictions of multivariate statistics. Figures 1 and 2 show an example for the distribution of variables. The first hypothesis to the distribution of the variables was a bimodal distribution of lognormally distributed data (Data > 0: skewed right, Data < 0: skewed left). The variables are transformed using $y = \text{sign}(x) \cdot \log(|x| + 1)$. The investigation of distributions leads to the finding of dichotomy in all six variables (positive or negative development). Scatter plots are used for a graphical display of the relationship between variables. The variables are also proofed by correlation coefficients. A strong correlation can not be detected.

Table 1 Overview of the dynamic variables

V	Label	Calculation
1	Population	Change in population as percentage, 1994–2004
2	Migration	(Move-in)–(move-out) as a number of persons, 1997–2004
3	Taxing capacity	Deviation to a tax-value (170 Euro/inh.) as percentage, 2003
4	Dwellings	Change in number of dwellings as percentage, 1994–2004
5	Employment rate	Change in employment as percentage, 1997–2004
6	Commuter ratio	(In-commuter)–(out-commuter)/population, 2004

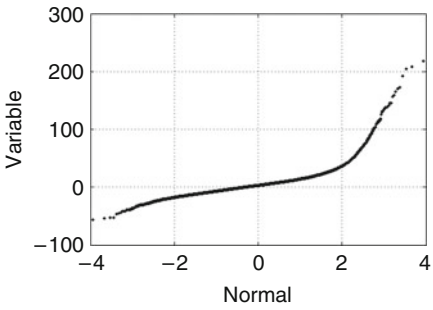


Fig. 1 QQ-Plot(population)

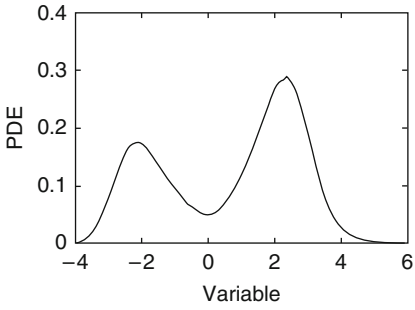


Fig. 2 PDE-Plot (Slogpopulation)

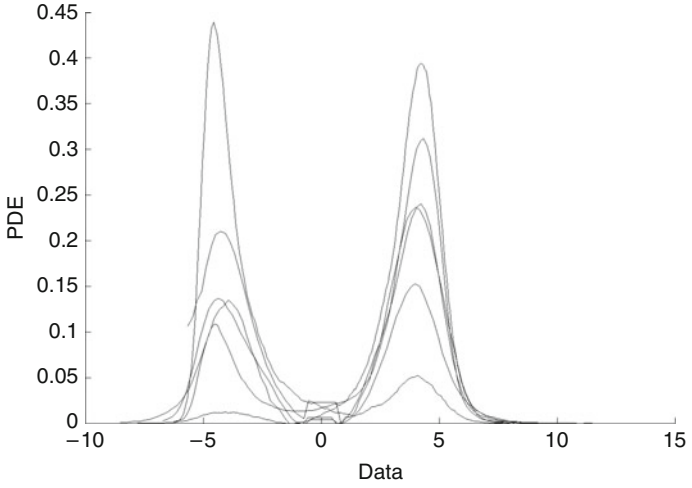


Fig. 3 Qualitative distribution of all variables using PDE

Figure 3 shows the qualitative properties of all transformed variables. Under the assumption that all combinations of characteristics exist 64 classes might describe the multidimensional dynamic of German communities. An equal distribution of objects to these classes gives a prior probability of $1/64 = 1.56\%$ for each class. Pertaining to the classification approach (e.g. U*-Matrix and subsequent

U*C-algorithm Ultsch 2005) and the properties of the Euclidian distance the data needs to be standardized.

3 Visualization and Clustering of Similar Dynamics

The power of self-organization allows the emergence of structure in data and supports its visualization, clustering and labeling concerning a combined distance and density based approach. To visualize high-dimensional data, a projection from the high dimensional space onto two dimensions is needed (= planar map). The map of an ESOM preserves the neighborhood relationships of the high dimensional data in a very good manner. The weight vectors of the neurons are thought as sampling point of the data. On the U*-Matrix a cluster structure in the dataset can be detected directly. Such visualization is used in tiled form to avoid border effects. The island view is realized by mask to reduce redundancies that means each neuron is nearly visible at once. The ESOM (50×150 neurons) is trained with the pre-processed dynamic data. The corresponding U*-Map (island view) delivers a geographical landscape of the input data on a projected map (imaginary axis). The cluster boundaries are expressed by mountains, which means the value of height is defining the distance between different objects, which are displayed on the z-Axis. A valley describes similar objects, characterized by small U-heights on the U*-Map. Data points found in coherent regions are assigned to one cluster. All local regions lying in the same cluster have the same dynamic properties. U*-Map of Fig. 4 includes the clustering results of the algorithm (U*C) and offers a visualization of hidden and unsuspected structures (13 cluster). Assigning a name to a cluster is one of the most important processes. In most times an aggregation process is necessary to build up a meaningful classification.

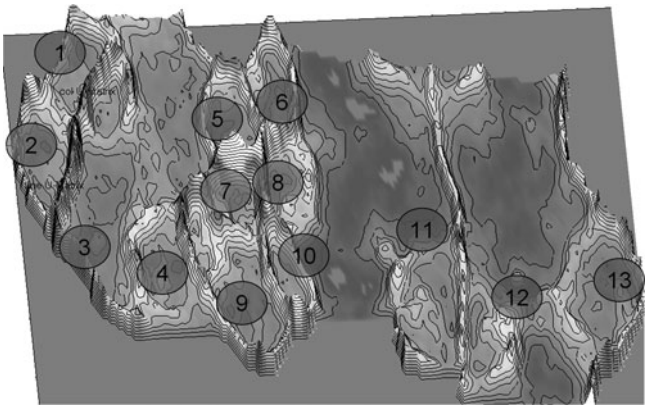


Fig. 4 Island of multidimensional community dynamics (U*-Map)

Table 2 Machine knowledge generation

Cluster with a positive employment rate
Cluster C1 (3377/174 communities): employment rate pos., migration pos. => subcluster C1.1/C1.2 (depending on population) Cluster C2 (678/127/310 communities): employment rate pos., migration neg. population pos. => subcluster C2.1 population neg. => subcluster C2.2/C2.3 (depending on commuter ratio)
Cluster with a negative employment rate
Cluster C3 (3197/324/276 communities): employment rate neg., migration pos. => subcluster C3.1/C3.2/C3.3 (depending on population and taxing-capacity) Cluster C4 (2114/158 communities): employment rate neg., migration neg. => subcluster C4.1/C4.2 (depending on dwellings) Cluster C5 (395 communities): employment rate, migration neg., population pos. Cluster C6 (1300 communities)

4 Explaining Patterns of Multidimensional Dynamics

The integration of Knowledge Discovery Techniques allows to understand the structure in a complementary form and supports the finding of an appropriate cluster aggregation and denomination. An extraction of explanations is realized and fits the minimal information for each cluster (Bishop 2006; Breiman et al. 1984). In this manner 2 or 3 variables are extracted to get a deeper view to the structure (Table 2).

5 Transition to Knowledge and Spatial Abstraction

Knowledge Conversion clarifies the understanding of detected structures and machine generated explanations. All results were validated mindful of the spatial analyst and yielded a spatial abstraction. In particular there are six main multidimensional dynamics and one class of outliers (Fig. 5). The localization of objects (Table 3) and additional spatial analysis were continuously used to proof the interpretation of cluster. The official city hierarchies (e.g. low-level/high-level-center), spatial typologies (e.g. central area or periphery), transport infrastructure (e.g. highway and railway system) and travel isochrones are used for a deeper interpretation. One sub-cluster should be highlighted ("Loser of the German reunification"). It contains many small rural communities and is explicit extracted in the Eastern part of Germany. The summarized communities are representing a dramatic development which is characterized by a massive negative job situation, a clear negative migration balance and a stationary or even decreasing number of dwellings. Furthermore the higher-order central places are often far away (50 ± 18 min).

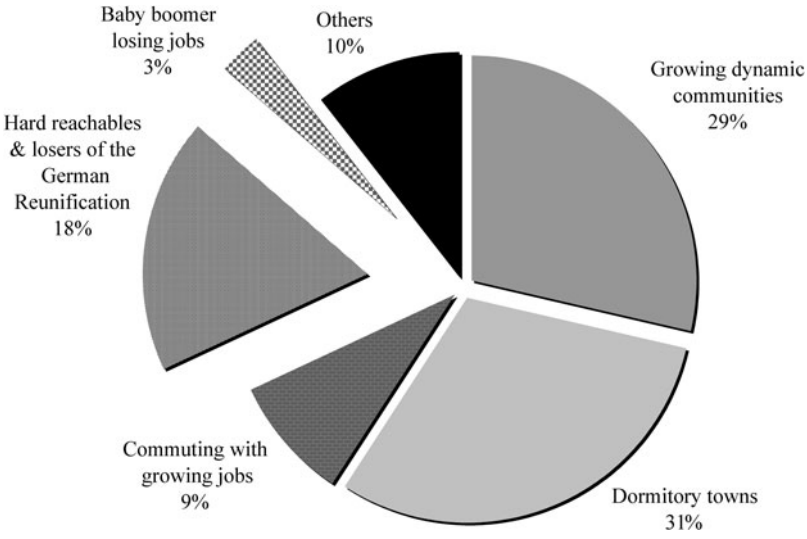


Fig. 5 Observed patterns of multidimensional dynamic

6 Discussion

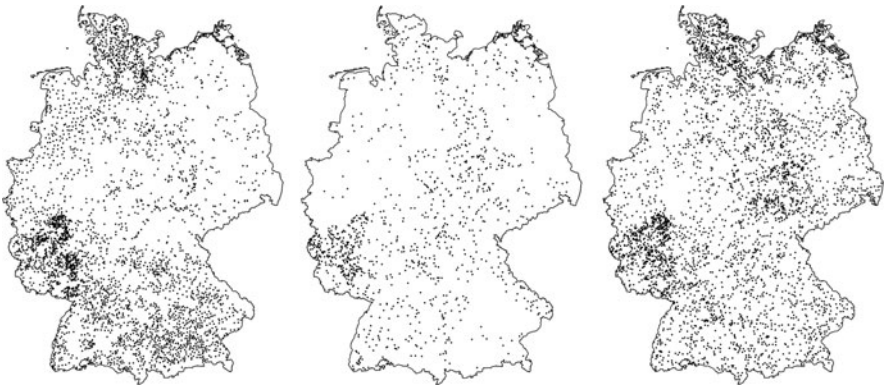
The aim of Knowledge Discovery implies the transition of data to knowledge. Such knowledge is required to be previously unknown, unsuspected and useful.

At first it was surprising that the observed distribution of communities was not equal in consideration of the prior probability ($1/64 = 1.56\%$). All 12,430 communities are distributed to just a few classes. Six main multidimensional dynamics are identified. Furthermore one class of outliers contains all classes in minority that means not fully occupied (below 1%). The approach poses some questions. Why do some classes not exist? What factors have an effect on the observed distribution of communities? The different size of communities has a strong influence on the concentration of cluster (e.g. Rheinland Pfalz). It might be good to take the modifiable unit problem into account (Openshaw 1984).

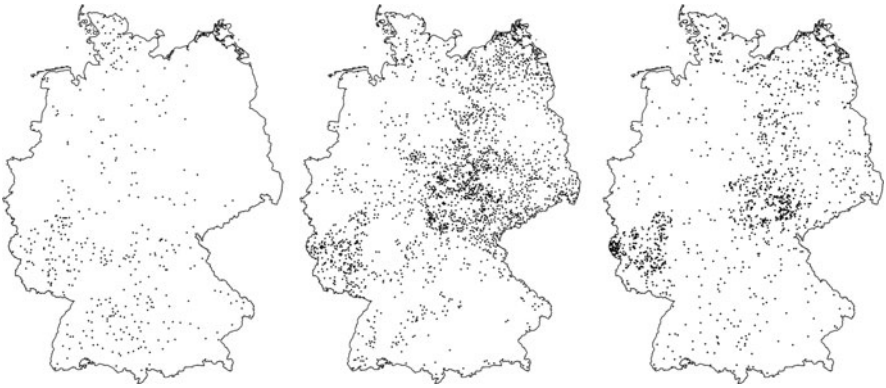
Secondly the dependencies between cluster and variables are discovered. According to the requirements of Knowledge Discovery such knowledge is previously unknown and due to the combined interpretation it might be utilized by decision makers within the field of spatial planning. The classes are addressed to the approved pressure factors for urban dynamic development (employment rate and population) and are extra represented by new and unsuspected combinations of dynamic properties.

Thirdly specific cluster should be investigated in detail by other structural and temporal parameters (e.g. age of population, buildings, infrastructure etc.). Due to the interpretation of cluster it is possible to proof several hypothesis about the German communities and their dynamic behaviour. These hypothesis are formed by the denomination of cluster (e.g. “Baby boomer losing jobs” or “Losers of the

Table 3 Localization of multidimensional dynamics



Growing Dynamic – Commuting with growing jobs – Dormitory towns



Baby boomer losing jobs – Hard reachables, losers of the reunification – others

German reunification"). Spatial outliers might be also of specific interest in the future (Shekhar et al. 2003).

7 Conclusion

The presented "Community Data Mining" approach provides the ability to identify hidden relationships and unusual patterns within a large amount of data (Behnisch 2009). An unsupervised classification approach is applied to 12,430 communities. The issue of the presented case study are shrinking and growing phenomena. In particular multidimensional patterns are explored to reveal knowledge about this spatiotemporal phenomenon.

First the pool of data is examined and the importance for the investigation of distributions is demonstrated according to the multidimensional dichotomy. One unsuspected result during the pre-processing is also the finding of a specific decision boundary in taxing capacity. It was therefore used for the optimized calculation of this variable (see Table 1) and might be interesting for further investigations. Afterwards it is shown that the use of Emergent SOMs is an appropriate method for clustering and classification. The advantage is to visualize the structure of data and later on to define a number of feasible cluster using U*C-algorithm while typical hierarchical algorithm often fail to examine complex structures (Ultsch 2006). The presented approach leads to the identification and abstraction of multidimensional dynamics. The structure control and interpretation was realized by finding a significant number of explanatory variables. Knowledge Conversion provides the transition from data to knowledge and it generates several hypotheses for further investigations. Six main dynamics of communities are therefore discovered.

In consideration of different multidimensional patterns it might be possible to think in new spatial relations and neighborhoods (e.g. comparative strengths, interregional communication and cooperation), thus, communities obtain a new urban condition. Common standards for a continuous observation of specific planning processes or mitigation measurements should be established based on multidimensional results. Especially the integration of temporary multidimensional investigations might encourage the short-term and long-term development of communities. Actually decision makers are not able to understand the underlying process that controls changes and patterns of changes. Furthermore procedures on the basis of knowledge-based systems are currently not sufficiently developed for a direct integration into the regional and urban planning and development processes. Such approaches might lead to a benchmark system for regional policy or to other strategic instruments such as semi or fully automated urban monitoring systems.

References

Behnisch, M. (2009). *Urban data mining*. Karlsruhe: Universitätsverlag Karlsruhe.

Bishop, C. (2006). *Pattern recognition and machine learning*. Berlin: Springer.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984). *Classification and regression trees*. California: Wadsworth.

Bundesamt, S. (Ed). (2005). *Qualitätsbericht Realsteuervergleich*. Wiesbaden: Statistisches Bundesamt (DESTATIS), Selbstverlag.

Gatzweiler, H. P., Meyer, K., & Milbert, A. (2003). Schrumpfende Städte in Deutschland? Ű Fakten und Trends. In BBR (Eds.), *Informationen zur Raumentwicklung* (Vol. 10/11, pp. 557–574). Bonn.

Openshaw, S. (1984). *The modifiable areal unit problem*. Norwich: Geo Books.

Siedentop, S., Kausch, S., Einig, K., Gössel, G. (2003). *Siedlungsstrukturelle Veränderungen im Umland der Agglomerationsräume.* Bundesamt für Bauwesen und Raumordnung (Ed.) (p. 114). Bonn.

Shekhar, S., Lu, C.-T., Zhang, P. (2003). A unified approach to detecting spatial outliers. In *GeoInformatica 7*(2), 139–166.

Ultsch, A. (2003). Pareto density estimation. In D. Baier & K. D. Wernecke (Eds.), *Innovations in classification, data science, and information systems* (pp. 91–100). Berlin: Springer.

Ultsch, A. (2005). U*C self-organized clustering with emergent feature map. In *Proceedings Lernen, Wissensentdeckung und Adaptivität* (pp. 240–246). Saarbrücken, Germany.

Ultsch, A. (2006). Analysis and practical results of U*C clustering. In R. Decker & H. J. Lenz (Eds.), *Advances in data analysis* (p. 6), Berlin: Springer.

Land Cover Detection with Unsupervised Clustering and Hierarchical Partitioning

Laura Poggio and Pierre Soille

Abstract An image segmentation technique relying on spatial clustering related to the single linkage approach has been put forward recently. This technique leads to a unique partition of the image domains into maximal segments satisfying a series of constraints related to local (α) and global (ω) intensity variation thresholds. The influence of such segmentation on clustering separability was assessed in this study, as well as the threshold values for segmentation maximising the cluster separability. The CLARA clustering method was used and the separability among clusters was calculated as the total separation between clusters. The clustering was applied to: (i) raw data; (ii) segmented data with varying α and ω parameters; and (iii) masked segmented data where the transition segments were excluded. The results show that the segmentation generally increases the separability of the clusters. The threshold parameters have an influence on the separability of clusters and maximising points could be identified while the transition segments were not completely included in one single cluster. The constrained connectivity paradigm could benefit land cover types/changes detection in the context of unsupervised object-oriented classification.

1 Introduction

The classification of a satellite image into land cover classes can be addressed at the level of pixels or segments generated by an image segmentation technique. The segmentation of an image can be defined as its partition into disjoint connected segments such that there exists a logical predicate returning true on each segment and false on any union of adjacent segments (Zucker 1976). Given an arbitrary logical predicate P, more than one valid segmentation may exist. If a unique segmentation is needed, the logical predicate P has to be based on equivalence relations (Jardine

L. Poggio (✉)
The Macaulay Land use Research Institute
e-mail: l.poggio@macaulay.ac.uk

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_49,

et al. 1967; Johnson 1967; Jardine and Sibson 1971). The segmentation of an image into meaningful regions can be achieved with numerous methods. Some examples of possible predicates are:

- Segments containing one and only one regional minimum. It leads to many possible segmentations.
- *Watershed transformation* (Vincent and Soille 1991): a steepest slope path linking each pixel of the segment to its corresponding minimum. It does not guarantee that there would be only one possible segmentation.
- *Iso-connectivity path* (Brice and Fennema 1970): two pixels are connected if and only if they can be joined by an iso-intensity path. This breaks the image into segments of uniform grey level. The method leads to a unique segmentation, but it normally gives a very fine partitioning with a lot of segments identified.
- *α-connectivity* (Nagao et al. 1979): two pixels of a grey image are connected if there exists a path of pixels linking these pixels and such that the grey level differences along adjacent pixels do not exceed a given threshold (α, local range parameter). This method is equivalent to single-linkage clustering (Gower and Ross 1969) and leads to an unique segmentation. However the resulting partition is often too coarse because some distinct objects are not identified when they are separated by one or more transitions with intensity steps $\leq \alpha$.
- *α,ω constrained connectivity* (Soille 2008) : the difference between the maximum and the minimum values of each connected component is limited with a second threshold value (ω, global range parameter) and a series of constraints related to α and ω is introduced in order to obtain a unique segmentation.

In this paper, the last segmentation method, α,ω constrained connectivity, is used. The image clustering often benefits from preliminary segmentation of the images. The aim of this work was to apply data clustering to the generated segments in order to (i) measure the influence of segmentation on clustering separability; (ii) evaluate which threshold values for segmentation maximise the cluster separability; and (iii) assess if the clustering could be helpful to detect segments corresponding to transition regions between adjacent segments.

2 Processing Flow

The processing flow of this study is briefly presented in Fig. 1. The test area is located in an agricultural region in France, south of Paris (Fig. 2). LANDSAT ETM+ data were used and bands 2,3,4 were selected in this preliminary approach. The bands were normalised in order to equalise the variance of the bands. The normalisation was done with the *znorm* transformation in which the mean of each attribute of the transformed set of data points is reduced to zero by subtracting the mean of each attribute from the values of the attributes and dividing the difference by the standard deviation of the attribute.

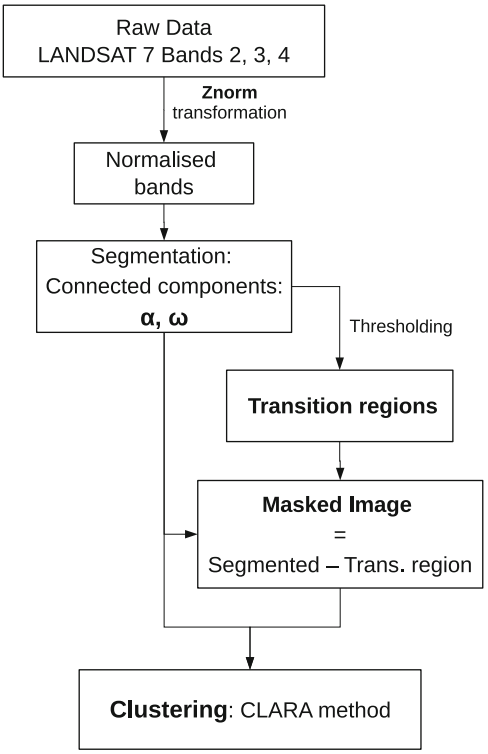


Fig. 1 Processing flow

(a) RGB representation of the sample data

(b) 4-3-2 false colour composite (Not simplified image in the text)

Fig. 2 Test data

3 Hierarchical Segmentation

The image was segmented with the method described in Soille (2008). This technique leads to a unique partition of the image domains into maximal segments satisfying a series of constraints related to local (α) and global (ω) intensity variations thresholds.

The (α, ω)-connected component of a pixel p was defined (Soille 2008) as the largest α_i-connected component of p such that (i) $\alpha_i \leq \alpha$ and (ii) its range (i.e. difference between its maximum and minimum values) is lower than or equal to ω:

$$(\alpha, \omega)\text{–CC}(p) = \bigvee \left\{ \alpha_i\text{–CC}(p) \middle| \alpha_i \leq \alpha \text{ and } R\Big(\alpha_i\text{–CC}(p)\Big) \leq \omega \right\}. \quad (1)$$

The existence of a largest α_i-connected component is secured due to the total order relation between the α_i-connected component of a pixel:

$$(\alpha, \omega)\text{–CC}(p) \subseteq (\alpha', \omega')\text{–CC}(p) \text{ for all } \alpha \leq \alpha' \text{ and } \omega \leq \omega'. \quad (2)$$

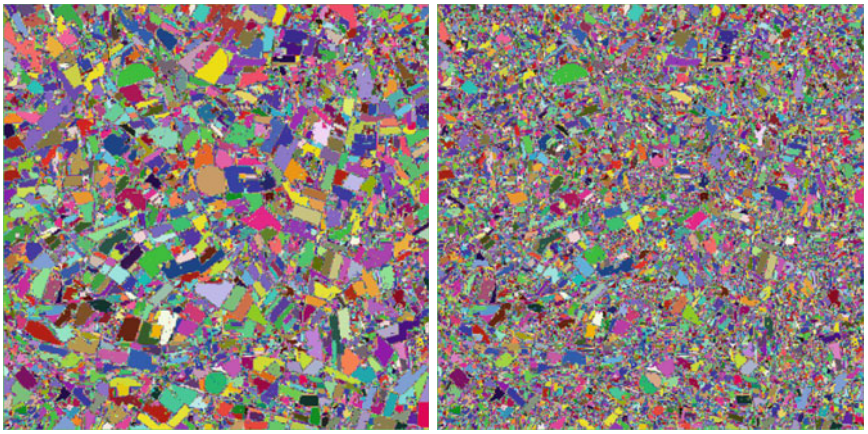
Finally two pixels p and q are said to be (α, ω)-connected if and only if $q \in (\alpha, \omega)\text{–CC}(p)$.

The increasing of α and ω values leads to a hierarchy of connected components. Partitions obtained with lower α-ω values are fully included into partitions obtained with larger α-ω values. The hierarchy of connected components can be graphically represented in a 3D dendrogram, where the leaves correspond to connected components of iso-intensity (Soille and Grazzini 2008).

The segmentation was done increasing α and ω values from 2 to 32 using various combinations of values. Figure 3 presents an example of labelling for different values of α and ω. Larger values create more homogeneous areas and thus a more simplified image (see Fig. 4). The simplified images were obtained setting each (α-ω)-connected component of Fig. 3 to the mean values for the considered bands.

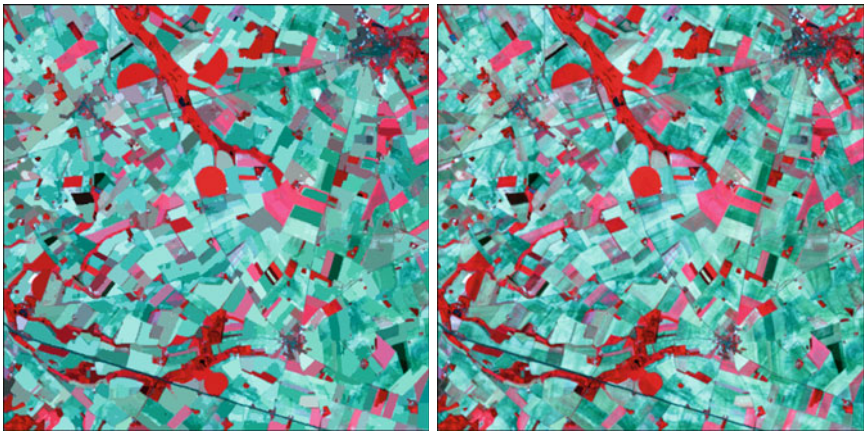
3.1 Transition Regions and Image Masking

The small regions created with the segmentation process were defined as regions that cannot contain the elementary structuring element defined by a pixel and its adjacent neighbours (Soille and Grazzini 2008). Such regions were subtracted from the original image to obtain masked images for each α, ω combination considered. Figure 5 shows that lower values of α and ω created a very high amount of small regions, corresponding to almost the whole image.



(a) $\alpha = \omega = 32$ **(b)** $\alpha = \omega = 8$

Fig. 3 Labelled $\alpha = \omega$ connected components

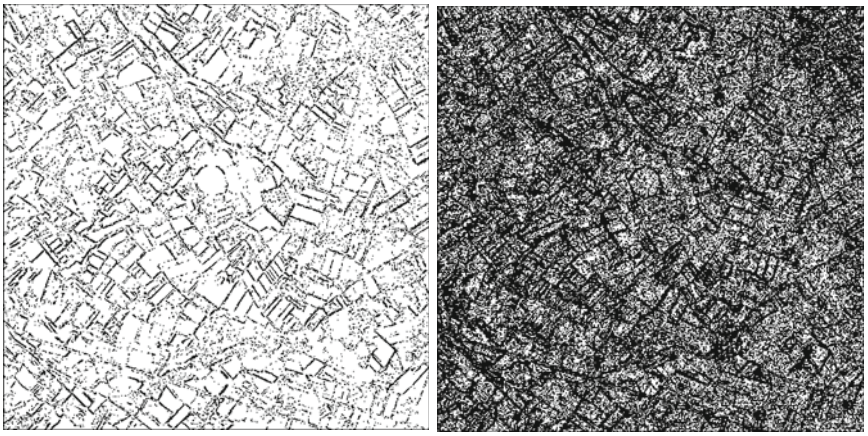


(a) $\alpha = \omega = 32$ **(b)** $\alpha = \omega = 8$

Fig. 4 Simplified images, obtained setting each (α-ω)-connected component to its band mean

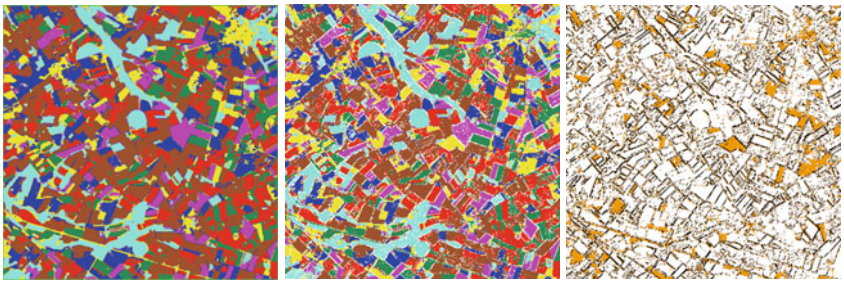
4 Unsupervised Clustering

The clustering method used was CLARA (Kaufman and Rousseeuw 1990) as implemented in the R software (http://www.r-project.org/index.html). The algorithm is based on the search for k representative objects (medoids) among the observations of the dataset. After finding a set of k medoids, k clusters are constructed by assigning each observation to the nearest medoid. The medoids are chosen to minimise the sum of the dissimilarities of the observations to their closest representative object. Compared to other partitioning methods, CLARA can deal with much larger



(a) $\alpha = \omega = 32$ **(b)** $\alpha = \omega = 8$

Fig. 5 Small regions identified by pixel size threshold from α-ω-connected components



(a) Not Masked **(b)** Masked **(c)** Differences

Fig. 6 Unsupervised classification on seven classes

datasets. This is achieved by considering sub-datasets of fixed size (sampsize) such that the time and storage requirements become linear rather than quadratic. Compared to the k-means approach, CLARA also accepts a dissimilarity matrix and it is more robust because it minimises a sum of dissimilarities instead of a sum of squared Euclidean distances.

The selection of the number of clusters was based on the Calinski criterion (Calinski and Harabasz 1974) obtained for a cascade of several partitions from a small to a large number of clusters. The resulting number was considered as best compromise among the different combinations of α and ω.

5 Classification and Cluster Separability

The classification results for masked and not masked images are presented in Fig. 6 for a relative high value of α and ω. The differences in classification occurred for the small regions, but also in the case of some larger objects. This suggests

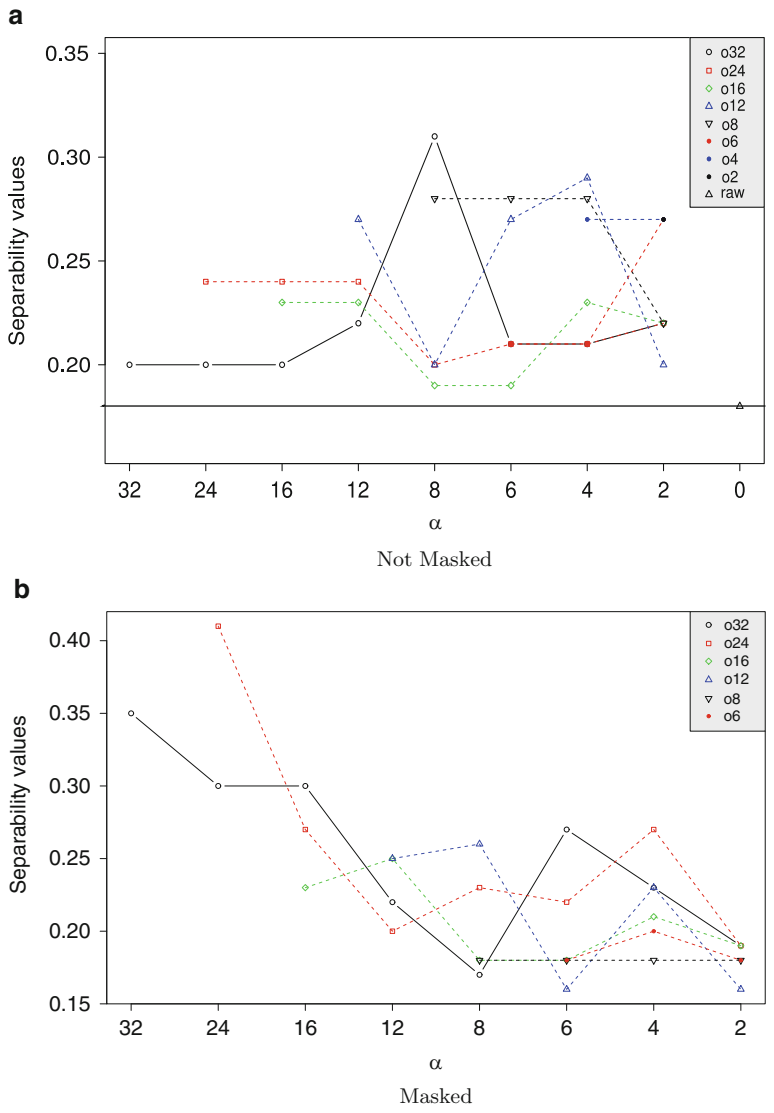


Fig. 7 Cluster separability measures

the importance of detection and correct classification of the small regions, as their influence on the classification of the image can be rather large.

A number of distance based statistics were computed for validation and comparison of clustering (Haldiki et al. 2001). The maximum separability was chosen as summarising parameter among masked and not-masked images for the various combinations of α and ω considered. Figure 7a presents the separability values for not-masked images. The values are rather scattered, but it is possible to identify a peak in the separability for values of $\omega = 32$ and $\alpha = 8$. The separability is

always increased compared to raw data ($\alpha = \omega = 0$). The separability values for the masked images with small regions excluded from the clustering are plotted in Fig. 7b. In this case it is possible to identify a trend with slightly increasing separability for increasing values of α and ω. The maximum separability is reached for $\alpha = \omega = 24$.

6 Concluding Remarks and Prospectives

The masking of the small regions proved to be useful to improve the unsupervised classification and to increase the cluster separability. The link with supervised classification is not yet fully explored. Further examples on different areas with different spectral information and more complex morphology are needed to get a potential pool of optimising values for α and ω as, indeed, any other connectivity constraint (see examples in Soille 2007). The constrained connectivity paradigm could benefit land cover types/changes detection in the context of unsupervised object-oriented classification.

References

Brice, C., & Fennema, C. (1970). Scene analysis using regions. *Artificial Intelligence, 1*(3), 205–226.

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3*, 1–27.

Gower, J., & Ross, G. (1969). Minimum spanning trees and single linkage cluster analysis. *Applied Statistics, 18*(1), 54–64.

Haldiki, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems, 17*, 107–145.

Jardine, C., Jardine, N., & Sibson, R. (1967). The structure and construction of taxonomic hierarchies. *Mathematical Biosciences, 1*(2), 173–179.

Jardine, N., & Sibson, R. (1971). *Mathematical Taxonomy*. London: Wiley.

Johnson, S. (1967). Hierarchical clustering schemes. *Psychometrika, 32*(3), 241–254.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis.* New York: Wiley.

Nagao, M., Matsuyama, T. T., & Ikeda, Y. (1979). Region extraction and shape analysis in aerial photographs. *Computer Graphics and Image Processing, 10*(3), 195–223.

Soille, P. (2007). On genuine connectivity relations based on logical predicates. In *14th International Conference on Image Analysis and processing, Proceedings* (pp. 487–492). 14th International Conference on Image Analysis and Processing, Modena, Italy, Sep 10–14.

Soille, P. (2008). Constrained connectivity for hierarchical image partitioning and simplification. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*(7), 1132–1145.

Soille, P., & Grazzini, J. (2008). Advances in constrained connectivity. Lecture Notes in Computer Science, *4992*, 423–433.

Vincent, L., & Soille, P. (1991). Watersheds in digital spaces – An efficient algorithm based on imemrsion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*(6), 583–598.

Zucker, S. (1976). Region growing: Childhood and adolescence. *Computer Graphics and Image Processing, 5*, 382–399.

Using Advanced Regression Models for Determining Optimal Soil Heterogeneity Indicators

Georg Ruß, Rudolf Kruse, Martin Schneider, and Peter Wagner

Abstract Nowadays in agriculture, with the advent of GPS-based vehicles and sensor-aided fertilization, large amounts of data are collected. With the importance of carrying out effective and sustainable agriculture getting more and more obvious, those data have to be turned into information – clearly a data analysis task.

Furthermore, there are novel soil sensors which might indicate a field's heterogeneity. Those sensors have to be evaluated and their potential usefulness should be assessed. Our approach consists of two stages, of which the first stage is presented in this article.

The data attributes will be comparable to the ones described in Ruß (2008). In the first stage, we will build and evaluate models for the given data sets. We will present a comparison between results using neural networks, regression trees and SVM regression. Results for an MLP neural network have been published in Ruß et al. (2008). In a future second stage, we will use the model information to evaluate and classify new sensor data. We will then assess their usefulness for the purpose of (yield) optimization.

1 Introduction

Due to the modernization and better affordability of state-of-the-art GPS technology and a multitude of available sensors, a farmer nowadays harvests not only crops but also growing amounts of data. These data are small-scale and precise – which is essentially why the combination of GPS, agriculture and data has been termed *precision agriculture*.

However, collecting large amounts of data often is both a blessing and a curse. There is a lot of data available containing information about a certain asset – here: soil and yield properties – which should be used to the farmer's advantage. This is a

G. Ruß (✉)
Otto-von-Guericke-Universität Magdeburg, Germany
e-mail: russ@iws.cs.uni-magdeburg.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_50,

common problem for which the term *data mining* or *data analysis* has been coined. Data analysis techniques aim at finding those patterns or information in the data that are both valuable and interesting to the farmer.

A common specific problem that occurs is yield prediction. As early into the growing season as possible, a farmer is interested in knowing how much yield he is about to expect. In the past, this yield prediction has usually relied on farmers' long-term experience for specific fields, crops and climate conditions. What if a computational model could be generated that allows to predict current year's yield based on past data and current year data? This problem of yield prediction encountered here is one which intelligent data analysis should be applied to. More specifically, multi-dimensional regression techniques could be used for yield prediction.

Nowadays, we can collect small-scale, precise data in-season using a multitude of sensors. These sensors essentially aim to measure a field's heterogeneity. In future work, these sensors will be assessed as to how useful they are for the purpose of yield prediction. For this work, this article should serve as an overview on the capabilities of different regression techniques used on agricultural yield data.

1.1 Research Target

The overall research target is to find those indicators of a field's heterogeneity which are suited best to be used for a yield prediction task. Since this should be done in-season, the sub-task here is one of multi-dimensional regression – predicting yield from past and in-season attributes. At a later stage, when multi-year data are available, models from past years could be used to predict present year's yield.

Therefore, this work aims at finding suitable data models that achieve a high accuracy and a high generality in terms of yield prediction capabilities. Since multi-year data are not yet available, the prediction can only be done in space using cross-validation, instead of in time. As soon as multi-year data are available, the models can be trained using these data for prediction in time. We will evaluate different types of regression techniques on different data sets. Since these models usually are strongly parameterized, an additional question is whether the model parameters can be carried over from one field to other fields which are comparable in (data set) size. This issue will also be addressed in this work. This is especially useful when new data have to be evaluated using one of the presented models.

1.2 Article Structure

Section 2 lays out the data sets that this work builds upon. The attributes and their properties will be presented shortly. Section 3 briefly presents four selected regression techniques from the data mining area which will be used for yield prediction.

Section 4 shows the results of the modeling/regression stage and provides answers to the aforementioned research questions.

At the end of this article, future work is pointed out and implementation details are provided.

2 Data Description

The data available in this work have been obtained in the years 2003–2006 on three fields near Köthen, north of Halle, Germany (GPS coordinates: Latitude N 51 40.430, Longitude E 11 58.110). All information available for these 65-, 72- and 32-hectare fields was interpolated using kriging (Stein 1999) to a grid with 10 by 10 meters grid cell sizes. Each grid cell represents a record with all available information. During the growing season of 2006, the latter field was subdivided into different strips, where various fertilization strategies were carried out. For an example of various managing strategies, see e.g. Schneider and Wagner (2006), which also shows the economic potential of PA technologies quite clearly. The fields grew winter wheat, where nitrogen fertilizer was distributed over three application times during the growing season.

Overall, for each field there are seven attributes – accompanied by the respective current year's yield (2004 or 2006) as the target attribute. Those attributes have been described in detail in Ruß et al. (2008), an overview is given below. In total, for the F04 field there are 5241 records, for F131 there are 2278 records, for F330 there are 4578 records, thereof none with missing values and none with outliers. In addition, a subset for F131 was available: in this subset, a special fertilization strategy was carried out which used a neural network for prediction and optimization – this data set is called F131net and has 1144 records.

In this work, data sets from three different fields are evaluated. A brief summary of two of the available data sets is given in Tables 1a and 1b. On each field, different fertilization strategies have been used. One of those strategies is based on a technique that uses a multi-layer perceptron (MLP) for prediction and optimization. This technique has been presented and evaluated in, e.g., Ruß et al. (2008); Ruß (2008) or Weigert (2006). For each field, one data set will contain all records, thus containing all the different fertilization strategies. In addition, a subset of F131 has been chosen to serve as a fourth data set to be evaluated.

3 Advanced Regression Techniques

As mentioned in the introduction, the task of yield prediction is essentially a task of multi-dimensional regression. Therefore, this section will serve as an overview about different regression techniques that are applicable to the yield data sets. We aim to evaluate these techniques on the data sets presented in the preceding section.

Table 1 Overview of the *F04* and *F131* data sets. The additional data sets *F330* and *F131net*, which is a subset of *F131*, are not shown as their statistical properties are very similar to those of *F04* and *F131*

(a) Data overview, F04					(b) Data overview, F131				
F04	*min*	*max*	*mean*	*std*	**F131**	*min*	*max*	*mean*	*std*
YIELD03	1.19	12.38	6.27	1.48	YIELD05	1.69	10.68	5.69	0.93
EM38	17.97	86.45	33.82	5.27	EM38	51.58	84.08	62.21	8.60
N1	0	100	57.7	13.5	N1	47.70	70	64.32	6.02
N2	0	100	39.9	16.4	N2	14.80	100	51.71	15.67
N3	0	100	38.5	15.3	N3	0	70	39.65	13.73
REIP32	721.1	727.2	725.7	0.64	REIP32	719.6	724.4	722.6	0.69
REIP49	722.4	729.6	728.1	0.65	REIP49	722.3	727.9	725.8	0.95
YIELD04	6.42	11.37	9.14	0.73	YIELD06	1.54	8.83	5.21	0.88

The regression task can be formalized as follows: the training set

$$T = \{\{x_1, \ldots, x_n\}, y_i\}_{i=1}^{N} \tag{1}$$

is considered for the training process, where $x_i, i = 1, \ldots, n$ are continuous input values and $y_i, i = 1 \ldots, N$ are continuous output values. Given this training set, the task of the regression techniques is to approximate the underlying function sufficiently well.

3.1 Introduction to Regression Techniques

Since one particular technique, namely MLPs, has been used successfully in previous work (Ruß et al. 2008; Ruß 2008), it is used as a reference model here. Three additional modeling techniques, namely RBF networks, regression trees, and support vector regression, will be presented, which are suitable for the task of yield prediction. The aforementioned techniques have, to the authors' knowledge, not been compared to each other when used with different data sets in the agriculture context. This section presents some of the background for each of the techniques before they will be evaluated in Sect. 4.

3.2 Neural Networks

In previous work multi-layer perceptrons (MLPs), a type of neural networks, have been used for a modeling task (Ruß et al. 2008; Ruß 2008) similar to the one encountered here. Furthermore, neural networks have shown to be quite effective in modeling yield of different crops (Drummond et al. 1998; Serele et al. 2000). The MLP model was established as a reference model against which further regression

techniques would have to compete. For a more detailed and formal description of MLP neural networks, it is referred to Hagan (1995) or Haykin (1998). The network layout and the parameters will be given in Sect. 4. In this work, the matlab implementation for the MLP network was used: `newff`.

Furthermore, a different type of neural network, a radial basis function (RBF) network, will be evaluated, since it is well-suited to the regression task. For this network, matlab's `newrb` function has been utilized.

3.3 Regression Tree

Regression as well as decision trees are usually constructed in a top-down, greedy search approach through the space of possible trees (Mitchell 1997). The basic algorithms for constructing such trees are CART (Breiman et al. 1984), ID3 (Quinlan 1986) and its successor C4.5 (Quinlan 1993). The idea here is to ask the question "which attribute should be tested at the top of the tree?" To answer this question, each attribute is evaluated to determine how well it is suited to split the data. The best attribute is selected and used as the test node. This procedure is repeated for the subtrees. For further information on the construction details and possible problems (such as overlearning) the reader is referred to Mitchell (1997). For this work the standard matlab implementation of `classregtree` has been utilized.

3.4 Support Vector Regression

Support Vector Machines (SVMs) are a supervised learning method discovered by Boser et al. (1992). However, the task here is regression, so the focus is on support vector regression (SVR). A more in-depth discussion can be found in Gunn (1998). Given the training set, the goal of SVR is to approximate a linear function $f(x) = \langle w, x \rangle + b$ with $w \in \mathbb{R}^N$ and $b \in \mathbb{R}$. This function minimizes an empirical risk function defined as

$$R_{emp} = \frac{1}{N} \sum_{i=1}^{N} L_\varepsilon(\hat{y} - f(x)), \tag{2}$$

where $L_\varepsilon(\hat{y} - f(x)) = \max((|\xi| - \varepsilon), 0)$. $|\xi|$ is the so-called slack variable, which has mainly been introduced to deal with otherwise infeasible constraints of the optimization problem, as has been mentioned in Smola and Schölkopf (1998). By using this variable, errors are basically ignored as long as they are smaller than a properly selected ε. L_ε is called ε-insensitive loss function. Other kinds of functions can be used, some of which are presented in Chap. 5 of Gunn (1998). To estimate $f(x)$, a quadratic problem must be solved. See Mejía-Guevara and Kuri-Morales

(2007) for the dual form of this problem. In this work, the SVMtorch implementation from Collobert et al. (2001) has been utilized. Its documentation also points out further details of the SVR process.

3.5 Linear Regression and Naive Estimator

For comparison reasons, two further prediction methods are employed to compare the advanced regression techniques against. The first of these is a simple multilinear regression estimator. The second is a naive estimator which simply reports the previous year's yield as the output yield of the current year.

3.6 Model Parameter Estimation

Each of the aforementioned four different models will be evaluated on the same data sets. One of the research goals here is to establish whether a model which has been used on one data set can be used on a different data set without changing its parameters. This would lead us to believe that comparable fields could use the same prediction model. Hence, the *F04* data set is used to determine the model parameters experimentally. Afterwards, the models are re-trained on the remaining data sets using the settings determined for *F04*. The parameter settings are given in Sect. 4.

For training the models, a cross-validation approach is taken. As mentioned in e.g. Hecht-Nielsen (1990), the data will be split randomly into a training set, a validation set and a test set. The model is trained using the training data and after each training iteration, the error on the validation data is computed. During training, this error usually declines towards a minimum. Beyond this minimum, the error rises – overlearning (or overfitting) occurs: the model fits the training data perfectly but does not generalize well. Hence, the model training is stopped when the error on the validation set starts rising. A size ratio of 8:1:1 for training, validation and test sets is used. The data sets are partitioned randomly 20 times and the models are trained. The models' performance will be determined using the root mean squared error (RMSE) and the mean absolute error (MAE) on the test set. It is assumed that the reader is familiar with these measures.

4 Regression Results

The models are run with the parameter settings given below. Those were determined experimentally on *F04* using a grid search, and carried over to the remaining data sets.

MLP A relatively small number of 10 hidden neurons is used and the network is trained until a minimum gradient of 0.001 is reached, using a learning rate of 0.25 and the *tangens hyperbolicus* sigmoid activation function.

Table 2 Results of running different models on different data sets. The best predictive model for each data set is marked in **bold** font

Model/Dataset	MAE				RMSE			
	F04	F131	F131net	F330	F04	F131	F131net	F330
MLP	0.3706	0.2468	0.2300	0.3576	0.4784	0.3278	0.3073	0.5020
RBF	0.3838	0.2466	0.2404	0.3356	0.5031	0.3318	0.3205	**0.4657**
REGTREE	0.4380	0.2823	0.2530	0.4151	0.5724	0.3886	0.3530	0.6014
SVR	**0.3446**	**0.2237**	**0.2082**	**0.3260**	**0.4508**	**0.3009**	**0.2743**	0.4746
LINREG	0.4285	0.3257	0.2766	0.3820	0.5578	0.4392	0.3871	0.5330
NAIVE	2.9061	0.6135	0.6492	4.7157	3.1253	0.7613	0.7847	4.8308

RBF For the radial basis function network, a radius of 1 is used for the radial basis neurons in the hidden layer. The algorithm, which incrementally adds neurons until the error goal of 0.001 is met, uses a maximum number of 70 neurons.

RegTree For the regression tree, the default settings of classregtree perform optimal; the full tree is pruned automatically and the minimum number of training examples below which no split should be done is 10.

SVR For the support vector regression model, the radial basis function kernel yields the best results, using the parameters $C = 60$, $\sigma = 4.0$ and $\xi = 0.2$.

Considering the results in Table 2, support vector regression obviously performs best on all but one of the data sets, regarding both error measures. Furthermore, SVR also is the model taking the least amount of computation time. Hence, the slight difference between the RMSE of SVR and RBF on the *F330* data set may be considered insignificant in practice when computational cost is also taken into account when deciding for a model. Regarding the understandability of the generated models, it would certainly be desirable to have the regression tree as the best model since simple decision rules can easily be generated from the tree. However, the regression tree performs worst in all of the cases. On the other hand, when comparing the hitherto reference model MLP with the current best model SVR, there is not much difference in the understandability of both models.

5 Conclusion

The results clearly show that support vector regression can serve as a better reference model for yield prediction than MLP. Even if the improvement should be statistically insignificant, the advantages of SVR over MLP remain. It is computationally less demanding, at least as understandable as the MLP and, most importantly, mostly produces better yield predictions. Furthermore, the comparison against a linear regression baseline and a naive estimator shows that the additional effort for using SVR is worth it.

Furthermore, the results also show that model parameters which have been established on one data set can be carried over to different (but similar with respect to the attributes) data sets. A model for identifying the most useful heterogeneity indicators is currently being evaluated.

5.1 Future Work

Due to the relatively high spatial resolution of the input data, the possible issue of spatial autocorrelation arises. This influences the modeling during the cross-validation stage. This will be investigated in future work.

Acknowledgements Experiments have been conducted using Matlab 2008a. The field trial data came from the experimental farm Görzig of Martin-Luther-University Halle-Wittenberg, Germany.

References

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory* (pp. 144–152). New York: ACM Press.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks.

Collobert, R., Bengio, S., & Williamson, C. (2001). Svmtorch: Support vector machines for large-scale regression problems. *Journal of Machine Learning Research, 1*, 143–160.

Drummond, S., Joshi, A., & Sudduth, K. A. (1998). Application of neural networks: precision farming. In *International Joint Conference on Neural Networks, IEEE World Congress on Computational Intelligence* (Vol. 1, pp. 211–215).

Gunn, S. R. (1998). *Support vector machines for classification and regression.* Technical Report, School of Electronics and Computer Science, University of Southampton, Southampton, U.K.

Hagan, M. T. (1995). *Neural network design (electrical engineering)*. Thomson Learning.

Haykin, S. (1998). *Neural networks: A Comprehensive Foundation* (2nd ed.). Englewood, Cliffs, NJ: Prentice Hall.

Hecht-Nielsen, R. (1990). *Neurocomputing*. Reading, MA, USA: Addison-Wesley.

Mejía-Guevara, I., & Kuri-Morales, A. (2007). Evolutionary feature and parameter selection in support vector regression. In *Lecture Notes in Computer Science* (Vol. 4827, pp. 399–408). Berlin, Heidelberg: Springer.

Mitchell, T. M. (1997). *Machine learning*. NY, USA: McGraw-Hill Science/Engineering/Math.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106.

Quinlan, R. J. (1993). *C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning)*. Los Altos, CA: Morgan Kaufmann.

Ruß, G., Kruse, R., Schneider, M., & Wagner, P. (2008). Estimation of neural network parameters for wheat yield prediction. In M. Bramer (Ed.), *Artificial Intelligence in Theory and Practice II* of *IFIP International Federation for Information Processing* (Vol. 276, pp. 109–118). Berlin: Springer.

Ruß, G., Kruse, R., Schneider, M., & Wagner, P. (2008). Optimizing wheat yield prediction using different topologies of neural networks. In J. L. Verdegay, M. Ojeda-Aciego, & Magdalena, L. (Eds.), *Proceedings of IPMU-08* (pp. 576–582). University of Málaga.

Ruß, G., Kruse, R., Wagner, P., & Schneider, M. (2008). Data mining with neural networks for wheat yield prediction. In P. Perner (Ed.), *Advances in Data Mining (Proc. ICDM 2008)* (pp. 47–56). Berlin, Heidelberg: Springer Verlag.

Schneider, M., & Wagner, P. (2006). Prerequisites for the adoption of new technologies – the example of precision agriculture. In *Agricultural Engineering for a Better World*, Düsseldorf: VDI Verlag GmbH.

Serele, C. Z., Gwyn, Q. H. J., Boisvert, J. B., Pattey, E., Mclaughlin, N., & Daoust, G. (2000). Corn yield prediction with artificial neural network trained using airborne remote sensing and

topographic data. In *2000 IEEE International Geoscience and Remote Sensing Symposium, 1*, 384–386.

Smola, A. J., & Schölkopf, B. (1998). *A tutorial on support vector regression.* Technical report, Statistics and Computing.

Stein, M. L. (1999). *Interpolation of Spatial Data : Some Theory for Kriging (Springer Series in Statistics)*. Berlin: Springer.

Weigert, G. (2006). *Data Mining und Wissensentdeckung im Precision Farming - Entwicklung von ökonomisch optimierten Entscheidungsregeln zur kleinräumigen Stickstoff-Ausbringung*. PhD thesis, TU München.

Local Analysis of SNP Data

Tina Müller, Julia Schiffner, Holger Schwender, Gero Szepannek, Claus Weihs, and Katja Ickstadt

Abstract SNP association studies investigate the relationship between complex diseases and one's genetic predisposition through Single Nucleotide Polymorphisms. The studies provide the analyst with a wealth of data and lots of challenges as the moderate to small risk changes are hard to detect and, moreover, the interest focusses not on the identification of single influential SNPs, but of (high-order) SNP interactions. Thus, the studies usually contain more variables than observations. An additional problem arises as there might be alternative ways of developing a disease.

To face the challenges of high dimension, interaction effects and local differences, we use associative classification and localised logistic regression to classify the observations into cases and controls. These methods contain great potential for the local analysis of SNP data as applications to both simulated and real-world whole-genome data show.

1 Introduction

The risk of developing a complex disease, e.g., cancer, is most likely not determined by a single factor, but rather influenced by several different external (e.g., lifestyle or environmental factors) and internal factors (e.g., genetic factors), cf. Fig. 1. Out of the possible genetic information sources, we focus on *SNPs* (*Single Nucleotide Polymorphisms*). A SNP refers to a single base exchange at a specific locus on the genome that is present in at least 1% of the population. For homologous chromosomes, there are three possible genotypes at each loci: the *homozygous reference* (if both chromosomes show the more frequent base), the *heterozygous* genotype (if one chromosome shows the more frequent and the other the less frequent base) and the *homozygous variant* (if both chromosomes show the less frequent base).

As interacting SNPs are assumed to influence the risk of developing diseases (Garte 2001), they might help to classify new observations into cases and controls.

T. Müller (✉)
Faculty of Statistics, TU Dortmund and SFB 475, Dortmund, Germany
e-mail: tmueller@statistik.tu-dortmund.de, tina.mueller@uni-dortmund.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_51,

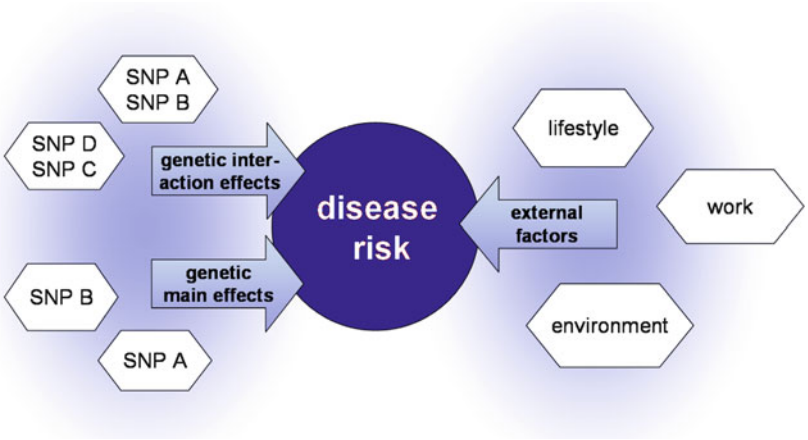


Fig. 1 Different factors and their interactions influence the disease risk

Unfortunately, lots of challenges accompany the analysis of SNP association studies. Since high throughput methods like SNP chips are widely used to generate the data, the analyst is usually confronted with more variables (up to several hundreds of thousands) than observations ($\sim$ 1,000), which results in the failure of standard procedures like logistic regression, especially if interactions are of interest. Furthermore, the effects on disease risk are assumed to be relatively small and might be due to interaction effects rather than to main effects alone. As can be seen in Fig. 1, instead of just one impact it is hypothesised that there are several competing ways of altering the disease risk (Clark et al. 2005). Therefore, we investigate local methods that can handle both interactions and large amounts of data. We will focus on *associative classification* (Liu et al. 1998; Müller et al. 2008) from the field of data mining and on *localised logistic regression* (Loader 1999; Tutz and Binder 2005; Schiffner et al. 2009), in comparison to ordinary *logistic regression* as well as *logic regression* (Ruczinski et al. 2003).

The two local methods will be introduced in Sect. 2, followed by a description of the data sets that will be analysed. The performance of all methods will be presented in Sect. 4. The final section gives a summary and an outlook on future work.

2 Methods

2.1 Associative Classification

A set of training data $(\mathbf{x}_n, y_n)$, $n = 1, \ldots, N$, is given with $\mathbf{x}_n \in I\!R^V$ indicating the genotype of an observation n on V loci and $y_n \in \{0, 1\}$ being the class label for the disease status. For applying associative classification, each SNP has to be

transformed into three dummy variables, each corresponding to one possible genotype. All $D = 3V$ dummy variables I_d, $d = 1, \ldots, D$, are now called *items*, while the observations are called *transactions* t_n. This yields the set of all transactions $\mathcal{T} = \{\sqcup_\infty, \sqcup_\in, \ldots, \sqcup_\mathcal{N}\}$ and the set of all items $\mathcal{I} = \{\mathcal{I}_\infty, \mathcal{I}_\in, \ldots, \mathcal{I}_\mathcal{D}\}$. A subset $s_j \subseteq \mathcal{I}$ is called an *itemset*.

An itemset s_j can have different characteristics, e.g., it can be frequent which is quantified by its *support* Sup_{s_j} (Borgelt and Kruse 2002):

$$Sup_{s_j} = \frac{1}{N} \sum_{n=1}^{N} \mathtt{I}_{\{s_j \subseteq t_n\}}, t_n \in \mathcal{T}, s_j \subseteq \mathcal{I},$$

with $\mathtt{I}_{\{\}}$ being the indicator function. If Sup_{s_j} exceeds a prespecified minimum support threshold, s_j is considered frequent. The support can be seen as an estimate for $Pr(s_j \subseteq t_n)$ for a random transaction t_n.

The concept of frequent itemsets can be extended to *association rules*. An association rule R_i consists of an *antecedent* (an itemset s_{R_i}) and a *consequent* (a single item I_{R_i}), giving information about how likely it is to observe the consequent in a transaction that is known to contain the antecedent. This likeliness can be quantified by the association rule's *confidence* $Conf_{R_i}$ (Borgelt and Kruse 2002):

$$Conf_{R_i} = \frac{\sum_{n=1}^{N} \mathtt{I}_{\{\{s_{R_i} \cup I_{R_i}\} \subseteq t_n\}}}{\sum_{n=1}^{N} \mathtt{I}_{\{s_{R_i} \subseteq t_n\}}}, t_n \in \mathcal{T}, s_{R_i} \subseteq \mathcal{I}.$$

An associations rule's support can be defined as the support of its antecedent, $Sup_{R_i} = Sup_{s_{R_i}}$. If both R_i's support and confidence exceed chosen thresholds, the respective rule is mined from the data.

The algorithm *apriori* (Agrawal 1993) searching for association rules is capable of handling huge amounts of data. As SNP data sets can easily comprise several hundred thousand variables, this feature is extremely helpful. The second advantage is the interpretability of association rules. If the consequent is restricted to consist of one of the two class labels only, the rule gives information about the likeliness of observing the respective disease status given the genetic profile of the antecedent. If the antecedent consists of more than one item, its itemset can be seen as an interaction between the SNPs involved.

For building a classifier, consider a set of mined association rules $\mathbf{R}$. A test observation $\mathbf{x}$ is classified based on the subset $\mathbf{R}(\mathbf{x})$ of $\mathbf{R}$ that is applicable (i.e., the antecedent of all $R_i \in \mathbf{R}(\mathbf{x})$ is part of $\mathbf{x}$). This ensures the locality, as every $\mathbf{x}$ defines its individual subset. Now, let $Co(R_i)$ be the consequent of association rule i (case: $Co(R_i) = 1$, control: $Co(R_i) = 0$). Test observation $\mathbf{x}$ is then classified according to the decision rule

$$\delta(\mathbf{R}(\mathbf{x})) := \begin{cases} 1, & \text{if } \frac{1}{|\mathbf{R}(\mathbf{x})|} \sum_{R_i \in \mathbf{R}(\mathbf{x})} Co(R_i) \geq \gamma \\ 0, & \text{else.} \end{cases} \tag{1}$$

In (1), γ gives the minimum fraction of all applicable rules that predict the status "case" in order to classify $\mathbf{x}$ as a case. We will use $\gamma = 0.100$ which has been a reasonable choice in preliminary analyses. The combination of association rules and classification in general is called associative classification (Liu et al. 1998), while this specific approach based on a vote of the rules will be called *AC Vote* in the following.

2.2 Localised Logistic Regression

The second local method we investigate is localised logistic regression (Tutz and Binder 2005).

Again, consider a set of training data $(\mathbf{x}_n, y_n)$, $n = 1, \ldots, N$. As y_n is a binary outcome, it cannot be modeled by, e.g., linear regression. Instead, a link function is needed to map the range of the dependent variable to $I\!R$. A famous choice is the logit link function, which results in the logistic regression model

$$\ln\left(\frac{\pi_n}{1-\pi_n}\right) = z_n'\beta, \quad n = 1, \ldots, N,$$

with $\pi_n = P(y_n = 1 \,|\, \mathbf{x}_n)$ being the class posterior probability, $z_n \in I\!R^Q$ being a design vector, e. g., $z_n' = (1, \mathbf{x}_n')$, and $\beta = (\beta_0, \ldots, \beta_{Q-1}) \in I\!R^Q$ being the parameter vector, to be estimated by maximising the log likelihood function

$$L(\beta) = \sum_{n=1}^{N} y_n \ln \pi_n + (1 - y_n) \ln(1 - \pi_n). \tag{2}$$

In ordinary logistic regression, a global model is fitted on the training data and a new observation $\mathbf{x}$ is classified according to the probability resulting by applying this model. In the local version, a separate model is fitted for each test observation, and weights $w_k(\mathbf{x}, \mathbf{x}_n)$ indicating the similarity of the new observation to the training observations $\mathbf{x}_n$ are introduced multiplicatively into the log likelihood function (Loader 1999):

$$L_{\mathbf{x}}^k(\beta) = \sum_{n=1}^{N} (y_n \ln \pi_n + (1 - y_n) \ln(1 - \pi_n)) \cdot w_k(\mathbf{x}, \mathbf{x}_n). \tag{3}$$

This ensures the influence of observations similar to $\mathbf{x}$ to be greater than for nonsimilar observations. The weighting function $w_k(\mathbf{x}, \mathbf{x}_n) = K\,(dist(\mathbf{x}, \mathbf{x}_n)/k)$ depends on a kernel $K(\cdot)$, a distance *dist* and a bandwidth k. We use

$$K(x) := \begin{cases} (1-|x|^3)^3, & \text{if } |x| \le 1 \\ 0, & \text{otherwise,} \end{cases}$$

(tricube kernel) and a distance based on flexible matching coefficients (Selinski and Ickstadt 2008) which are suitable for SNP data.

For each $\mathbf{x}$ the individual parameter estimate $\hat{\beta}_{\mathbf{x}}$ is computed by iterative Fisher scoring. In case of numerical problems due to local collinearities of predictors Tutz and Binder (2005) proposed to add the penalty term $-\lambda\beta' I\beta$ to $L^k_{\mathbf{x}}(\beta)$ in (3).

As in high dimensions local estimates are hardly local, we apply a variable selection (Tutz and Binder 2005) where the relevance of predictors is assessed by a local variant of Wald tests. After estimating $\hat{\beta}_{\mathbf{x}}$ the test statistic

$$c(\hat{\beta}_{\mathbf{x},\mathbf{q}}) = \frac{|\hat{\beta}_{\mathbf{x},\mathbf{q}}|}{\sqrt{\widehat{Var}(\hat{\beta}_{\mathbf{x},\mathbf{q}})}}, \quad q = 1, \ldots, (Q-1) \tag{4}$$

is calculated and predictors exceeding a threshold c_β are selected. The variance of the parameter estimates in (4) is estimated by the inverse Fisher matrix. The entire process of classifying a test observation $\mathbf{x}$ is given by:

1. Calculate weights $w_k(\mathbf{x}, \mathbf{x}_n)$.
2. Determine $\hat{\beta}_{\mathbf{x}}$ by iterative Fisher scoring (possibly with penalty λ).
3. If the Fisher scoring converges, select predictors, recalculate the weights, and repeat the Fisher scoring for the selected influential factors.
4. Use the (reduced) model to predict the class for $\mathbf{x}$.

This procedure depends on the three parameters k, λ, and c_β that can be optimized by means of a grid search.

3 Data

To compare the performance of the classification methods, we will analyse a simulation study as well as a genome-wide real-world data set.

All **simulated data** sets (using the software SNaP Nothnagel 2002) contain 500 cases and 500 controls with categorical values for 40 SNP variables. We designed four scenarios in which interactions between different numbers of SNPs with given penetrances (probability of developing the disease given a certain genotype of the causative SNPs) influence the disease status. In the first scenario, one interaction between two SNPs (one two-way interaction) determines the disease status. In scenarios 2 and 3, two and three mutual independent two-way interactions, respectively, are responsible for the outcome disease. Finally, two independent three-way interactions influence the disease risk in scenario 4. We chose minor allele frequencies between 0.1 and 0.3 for the causative SNPs (consistent with real-world studies) and penetrances rising with more variants within the SNP interaction. There are ten data sets in each scenario, each used once as training data and once as test data for classification.

Our real-world data set is a subset of the **HapMap data** (The International HapMap Consortium 2003) comprising 45 unrelated Han Chinese from Beijing and 45 unrelated Japanese from Tokyo. Thus, ethnicity is used as a class label. The subset consists of 157 SNPs showing the largest values of Pearson's χ^2-statistics amongst the 121774 SNPs from the Nsp array of the Affymetrix GeneChip Mapping 500K Array Set that express all three genotypes and have a minor allele frequency greater than 0.1.

4 Results

All results were obtained using the software package R 2.8.1 (R Development Core Team 2008), in particular packages `arules` (Hahsler 2007) and `LogicReg` (Kooperberg and Ruczinski 2008). AC Vote and localised logistic regression (LLR) were applied to all data sets and compared to the results of ordinary logistic regression (with main effects only) and logic regression (Ruczinski et al. 2003), a method especially designed for SNP data. While the local methods have two advantages over logistic regression (they present interactions and give an individual result for every test observation) and the advantage of individual results over logic regression, we compare all classification performances by assessing the misclassification rate (MCR).

As can be seen in Fig. 2, for the first simulated scenario (one causative two-way interaction) all MCRs are between 0.210 and 0.252, with LLR leading to the lowest MCR. In the second scenario, all MCRs are quite similar and have a maximum

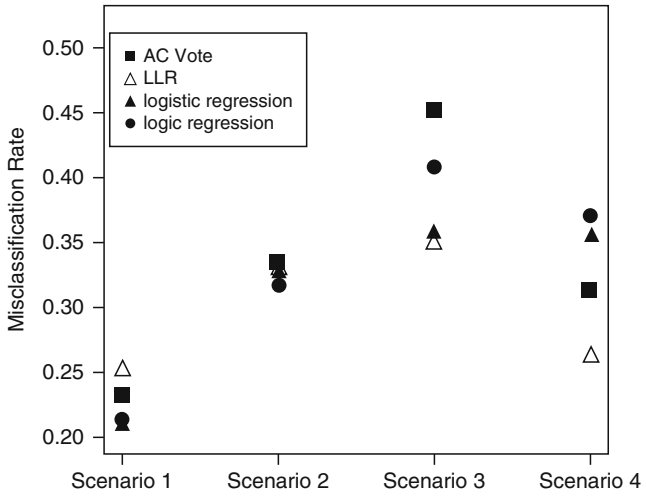


Fig. 2 Misclassification rates achieved by AC Vote, LLR, logic regression and logistic regression for the simulation study

Table 1 Misclassification rates for the HapMap data achieved by logistic regression, LLR, logic regression and AC Vote.

	logistic regression	LLR	logic regression	AC Vote
MCR HapMap	—	0.500	0.144	0.011

at 0.335. Three causative two-way interactions (Scenario 3) seem to be the hardest classification task. AC Vote yields the highest MCR (0.452), mainly due to lack of sensitivity. As three interactions induce many fragmental association rules (containing only part of an interaction) that can falsely predict control status, the whole performance of AC Vote gets worse. The approaches based on logistic regression still do fairly well (0.357). In Scenario 4, AC Vote and especially logistic regression outperform the two other methods. The success of logistic regression, similarly to the third scenario, is probably due to the simulation process (cases in Scenario 4 can mostly be determined by one of the interacting SNPs), therefore logistic regression is able to pick up the signal even if it does not necessarily trace it back to the causative interaction. AC Vote gives good results if the interaction of interest is present in the applicable rule set, averaging to a reasonable MCR.

Logistic regression fails in classifying the HapMap example as it contains more variables than observations. LLR is still feasible (with $\lambda = 0.700$), but yields no reasonable result. The data set is well distinguishable into Han Chinese and Japanese, with AC Vote yielding a MCR of only 0.011 (cf. Table 1).

5 Summary and Discussion

We showed on different SNP data sets that local methods, in particular associative classification and localised logistic regression, are applicable and yield good and interpretable results. Their misclassification rates are in some cases best, but not always better than standard methods (logistic regression and logic regression). However, they keep the advantage of locality (in contrast to logistic and logic regression) and of allowing for interaction effects in high dimension (in contrast to logistic regression).

In this paper we analysed only a subset of the HapMap data that was already sufficient for a good classification of the observations. If we use the genome-wide data, however, AC Vote is the only method among the four, and among very few that can detect high order SNP interactions that is feasible without an initial reduction of the data to a couple of hundreds of SNPs.

Still, there is lots of room for improvement, both on computational aspects and on sophisticated threshold and parameter choices. E.g., the voting scheme of AC Vote, saying that a new observation will be classified as a case if at least 10% of all applicable rules predict this outcome, can be refined into an approach where each rule gets an individual weight in the voting process to improve the misclassification rate.

A further adaption of the computational aspects of localised logistic regression will be investigated to allow for more input variables.

Acknowledgements Financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") is gratefully acknowledged.

References

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In P. Buneman, and S. Jajodia (Eds.), *Proceedings of 1993 ACM SIGMOD International Conference on Management of Data* (pp. 207–216). Washington, DC, May 26–28.

Borgelt, C., & Kruse, R. (2002). Induction of association rules: Apriori implementation. In W. Haerdle, and B. Roenz (Eds.), *COMPSTAT* (pp. 395–400).

Clark, A. G., Boerwinkle, E., Hixson, J., Sing, C. F. (2005). Determinants of the success of whole genome association testing. *Genome Research, 15*, 1463–1467.

Garte, S. (2001). Metabolic susceptibility genes as cancer risk factors: Time for a reassessment? *Cancer Epidemiology Biomarkers & Prevention, 10*, 1233–1237.

Hahsler, M., Gruen, B., & Hornik, K. (2007). arules: Mining association rules and frequent itemsets, R package version 0.6-3.

The International HapMap Consortium. (2003). The international HapMap project. *Nature, 426*.

Kooperberg, C., & Ruczinski, I. (2008). LogicReg: Logic Regression, R package version 1.4.8.

Liu, B., Hsu, W., & Ma, Y. (1998). Integrating classification and association rule mining. In R. Agrawal and P. Stolorz (Eds.), *Proceedings of 4th International Conference on Knowledge Discovery and Data Mining* (pp. 80–86).

Loader, C. (1999). *Local regression and likelihood*. Springer Series in Statistics and Computing. New York: Springer.

Müller, T., Schwender, H., & Ickstadt, K. (2008). Finding SNP interactions. In M. Ahdesmäki, K. Strimmer, N. Radde, J. Rahnenführer, K. Klemm, H. Lähdesmäki and O. Yli-Harja (Eds.), *Proceedings of 5th International Workshop on Computational Systems Biology (WCSB)* (pp. 109–112).

Nothnagel, M. (2002). Simulation of LD block-structured SNP haplotype data and its use for the analysis of case-control data by supervised learning methods. *American Journal of Human Genetics, 71*(Suppl.)(4), A2363.

R Development Core Team. (2008). *R: A Language and environment for statistical computing. R Foundation for Statistical Computing.* Austria: Vienna.

Ruczinski, I., Kooperberg, C., & LeBlanc, M. (2003). Logic regression. *Journal of Computational and Graphical Statistics, 12*, 475–511.

Schiffner, J., Szepannek, G., Monthé, T., & Weihs, C. (2009). Localized logistic regression for categorical influential factors. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker (Eds.), *Data Analysis, Machine Learning and Applications* (pp. 69–76). Heidelberg: Springer.

Selinski, S., & Ickstadt, K. (2008). Cluster analysis of genetic and epidemiological data in molecular epidemiology. *Journal of Toxicology and Environmental Health A, 71*, 835–844.

Tutz, G., & Binder, H. (2005). Localized classification. *Statistics and Computing, 15*, 155–166.

Airborne Particulate Matter and Adverse Health Events: Robust Estimation of Timescale Effects

Massimo Bilancia and Francesco Campobasso

Abstract Numerous epidemiological studies based on time-series analysis have shown associations between morbidity/mortality caused by respiratory and cardiovascular adverse events and chronic exposure to airborne particles (Bell et al. 2004), but a considerable uncertainty remains to be seen. This begs the question of whether these associations represent premature morbidity within only a few days among those people already near to acute health events. Statistical aspects of such a displacement effect (or *harvesting*) have been discussed by several authors (Dominici et al. 2003 and references therein); a reasonable underlying hypothesis is that mortality/morbidity displacement is associated with shorter timescales, while longer time scales are supposed to be resistant to displacement. If associations reflect only harvesting, the effect of air pollution on morbidity can be considered as having a limited impact from a public-health point of view. In this paper we discuss a new approach to assess the effect of short term changes in air pollution on acute health effects. Our method is based on a Singular Spectrum Analysis (SSA) decomposition of airborne particulate matter time series into a set of exposure variable, each one representing a different timescale. An advantage of our approach is that timescales need not to be set prior to their estimation.

1 Introduction

Numerous epidemiological studies based on time-series analysis have shown associations between morbidity/mortality caused by respiratory and cardiovascular adverse events and chronic exposure to airborne particles (Bell et al. 2004). Particles with an aerodynamic diameter of less than 10 microns are referred to as PM_{10}; they may be inhaled reaching upper airways and lungs, with risk for health. Despite this growing body of evidence, a considerable uncertainty remains to be seen; this begs

M. Bilancia (✉)
Dipartimento di Scienze Statistiche "Carlo Cecchi", Università degli Studi di Bari, Italy
e-mail: mabil@dss.uniba.it

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_52,

the question of whether these associations represent premature morbidity within only a few days among those already near to acute health events. Such a displacement (or *harvesting*) effect has been discussed by several authors (see, for example, Dominici et al. 2003 and references therein). A reasonable underlying hypothesis is that mortality/morbidity displacement is associated with shorter timescales, while longer time scales are supposed to be resistant to mortality displacement.

A prominent approach to timescale effect estimation was introduced in Dominici et al. (2003), in which the authors developed a methodology based on the Discrete Fourier Transform (DFT) by partitioning the base interval $[0, \pi]$ into a given set of Fourier frequencies, to obtain a decomposition of pollutant series into a set of orthogonal predictors, each one representing a different timescale. A drawback of such methodology is that timescales need to be set prior to estimation; this is why we introduce an alternative approach based on Singular Spectrum Analysis (SSA, see Golyandina et al. 2001), which can be defined as a model-free approach to decompose time series into easy-to-interpret components (such as trend, mid and short-period waveforms) without any prior knowledge of relevant timescales. On the basis of real data, we contrast our approach with results obtained by Fourier decomposition.

2 Materials and Methods

2.1 Data and Statistical Approach to Estimation of Associations at Different Timescales

Our case study is based on daily measurements of PM_{10} obtained in Bari (Apulia, Italy); particulate matter and meteorological variables measurements were obtained from the city monitoring network maintained by the Municipality of Bari (Department of Environmental Protection and Health). Original pollutant data were obtained on a bi-hourly basis; details about pre-processing and outlier filtering are described elsewhere (Bilancia and Stea 2008). Epidemiological data were obtained from the Apulian Regional Epidemiological Center; we used the daily time-series of hospitalized people among residents in the city of Bari (in total $N = 579$ days between June 1th, 2000 and December 31th, 2001), diagnosed as suffering from pulmonary diseases (ICD-IX Classification: 460–519). Data are shown in Fig. 1 below.

We assume that following over-dispersed Generalized Additive Model (GAM) holds for daily counts Y_t of adverse health events

$$Y_t \,|\mu_t \overset{ind.}{\sim} \text{Poisson}\,(\mu_t) \quad \text{Var}\,(Y_t) = \phi\mu_t \tag{1}$$

with $\phi > 1$ and (see Dominici et al. 2003 for a wide review of related statistical methods)

$$\log(\mu_t) = \alpha + \beta x_t + \text{DOW}_t + \mathcal{S}(t, \delta_1) + \mathcal{S}(\text{temp}_t, \delta_2) + \mathcal{S}(\text{umr}_t, \delta_3) \tag{2}$$

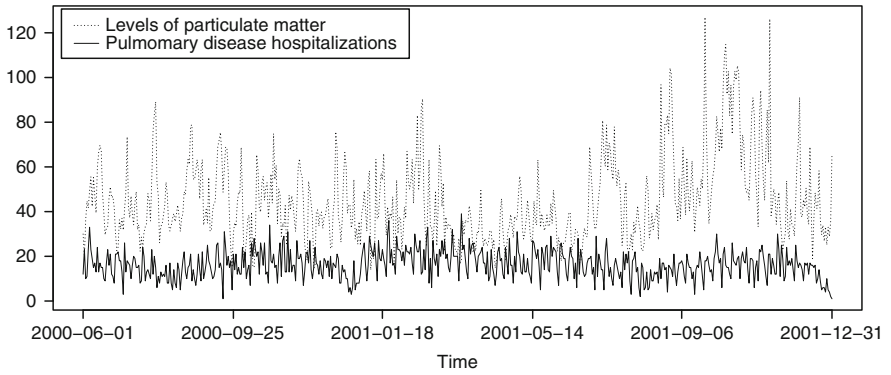


Fig. 1 Daily time series of hospitalizations for pulmonary diseases (ICD-IX Classification: 460–519), and levels of particulate matter with an aerodynamic diameter less than 10 $\mu g/m^3$ (PM_{10}) for Bari, Apulia, Italy, during the period June 1th, 2000 – December 31th, 2001

where x_t is the daily PM_{10} concentration measured in $\mu g/m^3$, DOW_t is a six-dimensional vector of dummy variables modeling day-of-the-week effects and $\mathcal{S}(t, \delta_1)$ is a smooth term function of calendar time protecting for confounding by seasonality and longer-term trends (the degree of roughness being controlled by the smoothing parameter δ_1). Further smooth confounders entering the model are the daily temperature ($temp_t$) measured in °C and the daily relative humidity (umr_t) expressed as percentage (meteorological variables may affect the pollution-morbidity association, Samet et al. 1998).

In our approach the term βx_t is replaced with the linear function $\sum_{\ell=1}^{p} \beta_\ell x_{\ell t}$, assuming that the following linear decomposition holds

$$x_t = \sum_{\ell=1}^{p} x_{\ell t} \tag{3}$$

where $x_{1\cdot}, \ldots, x_{p\cdot}$ is a set of suitable predictors each one representing a different timescale. The model (3) estimates the effect of airborne pollution on health at different timescales; for given timescale ℓ, this effect is quantified by the *adjusted relative risk* (ARR) $\exp(10 * \beta_\ell)$, which represents the ratio of risks between subpopulations defined by $x_{\ell\cdot} = x^{\star}_{\ell\cdot}$ and $x_{\ell\cdot} = x^{\star}_{\ell\cdot} + 10$ (this ratio does not depend on the reference level $x^{\star}_{\ell\cdot}$).

Automatic model selection is a suitable way to set the degrees of freedom associated with smooth terms entering model (2). In our case, given that the scale parameter ϕ is unknown, smoothing parameter estimation can be based on the mean square prediction error, that is the average squared error in predicting a new observation using the fitted model. Generalized Cross Validation (GCV) is a computationally feasible approach to expected prediction error estimation (Wood 2004). As in the present paper, several competing linear decompositions (3) are indeed possible. For this reason, GCV scores can be used as well to assess the predictive

accuracy of corresponding models, which specifically differ from each other for the predictors representing timescales.

2.2 Fourier Decomposition

We begin by summarizing the Fourier decomposition proposed in Dominici et al. (2003). Suppose that the time series $\{x_t : t = 1, \ldots, N\}$ has length N and let $\omega_j = 2\pi j/N$ be the j-th Fourier frequency for $\omega_j \in [0, \pi]$. For $j = 1$ we have the first harmonic, whose angular frequency and period are respectively $\omega_1 = 2\pi/N$ and $T_1 = 2\pi/\omega_1 = N$. Such a harmonic is a one-cycle in the length of data waveform which describes the longest-term fluctuation. In the case N is even the shortest-term fluctuation corresponds to $j = N/2$, which gives $\omega_{N/2} = \pi$ and $T_{N/2} = 2$, whereas, if N is odd, $j = (N-1)/2$ has to be set. Consider now the Discrete Fourier Transform (DFT) of the data

$$d(\omega_j) = \frac{1}{N} \sum_{t=1}^{N} x_t \exp(-\imath \omega_j t) \tag{4}$$

for $0 \leq j < N - 1$ so that $0 \leq \omega_j < 2\pi$. Apparently redundant specification of frequencies ω_j in the range $[0, 2\pi[$ disappears when we note that for $j > N/2$, that is $\omega_j \in]\pi, 2\pi[$, we have $d(\omega_j) = \overline{d(\omega_{N-j})}$, where $\overline{d(\cdot)}$ denotes the complex conjugate of $d(\cdot)$. Consider now a partition of the interval $[0, \pi]$ based on p Fourier frequencies, that need to be set prior to computations, say $\left[\omega_0, \omega_1, \ldots, \omega_\ell, \ldots, \omega_p, \omega_{p+1}\right]$ with $\omega_0 = 0$ and $\omega_{p+1} = \pi$, and let $I_\ell =]\omega_{\ell-1}, \omega_\ell] \cup [\omega_{N-\ell}, \omega_{N-\ell+1}[$. A linear decomposition (3) follows by suitable rearrangement of the inverse DFT terms

$$x_t = \sum_{j=0}^{N-1} d(\omega_j) \exp(\imath \omega_j t) = \sum_{\ell=1}^{p+1} \left[\sum_{\omega_j \in I_\ell} d(\omega_j) \exp(\imath \omega_j t) \right] = \sum_{\ell=1}^{p+1} x_{\ell t} \tag{5}$$

2.3 Singular Spectrum Analysis

In this section we review the basics of SSA and propose a refined version of the algorithm described in Bilancia and Stea (2008) (a detailed exposition of SSA can be found in Golyandina et al. 2001). Consider again a realization of a one-dimensional time series $\{x_t : t = 1, \ldots, N\}$ and let L be a fixed integer, called the *window length*, with $1 < L < N/2$. The *embedding* procedure consists in defining a sequence of $K = N - L + 1$ *lagged vectors* $X_i^{(L)} = (x_i, x_{i+1}, \ldots, x_{i+L-1})^T$ and the *trajectory matrix* given by

$$X = \left[x_{ij}\right]_{i,j=1}^{L,K} = \left[X_1^{(L)}, \ldots, X_K^{(L)}\right] \tag{6}$$

The embedding procedure is closely linked to the method of delays in dynamical system theory; it is worth noting that the matrix X is a Hankel matrix, i.e. all the elements x_{ij} along the secondary diagonals such that $i + k = \text{constant}$ are equal and its columns are copies of overlapping segments of time series. Vice versa, if any rectangular matrix X is Hankel, then X it is the trajectory matrix of some time series.

The second phase of the algorithm includes the computation of the Singular Valued Decomposition (SVD) of the matrix X: let $S = XX^T$, $d = \text{rank}(S)$, λ_i the eigenvalues of S in decreasing order, U_i the corresponding eigenvectors and $V_i = X^T U_i / \sqrt{\lambda_i}$ for $i = 1, \ldots, d$. The SVD of the trajectory matrix X is given by

$$X = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^T \tag{7}$$

The squared roots $\sqrt{\lambda_i}$ are known as *singular values*, while U_i and V_i are respectively called the *left* and *right singular vectors*. Finally the collection $\left(\sqrt{\lambda_i}, U_i, V_i\right)$ is called the *i-th eigentriple* of the matrix X. Hence, the trajectory matrix is decomposed into a sum of elementary rank-one, pairwise bi-orthogonal matrices. It can be also proved that $\sum_{i=1}^{d} \lambda_i$ equals the squared Frobenius-Perron norm of the matrix X, as well as that λ_i is the squared Frobenius-Perron norm of the component matrix $X_i = \sqrt{\lambda_i} U_i V_i^T$ $(i = 1, \ldots, d)$. Thus the ratio $\sum_{i=1}^{r} \lambda_i / \sum_{i=1}^{d} \lambda_1$ measures the degree of approximation of the trajectory matrix, when X is approximated by the sum of the first r terms in the left-hand side of (7).

The grouping phase is to select p disjoint subsets from the index set $\{1, \ldots, d\}$, say $\left\{I_1, \ldots, I_p\right\}$, with $I_j = \left(j_1, \ldots, j_{n_j}\right)$ such that the SVD decomposition (7) can be formulated as $X = X_{I_1} + \cdots + X_{I_p}$ with $X_{I_j} = X_{j_1} + \cdots + X_{j_{n_j}}$ $(j = 1, \ldots, p)$. Suppose now that each matrix X_{I_j} is Hankel; hence they are trajectory matrices from which component series can be reconstructed. Alternatively *diagonal averaging* can be applied (Buchstaber 1994), which is the result of the application of a suitable orthogonal linear projection operator $\mathcal{H}$ of minimum norm to both sides of the decomposition $X = X_{I_1} + \cdots + X_{I_p}$. It is easily proved that $\mathcal{H}X = X$ and that $\mathcal{H}X_{I_j}$ is Hankel, from which the linear decomposition (3) of the original series into p *reconstructed components* (RC) can be easily recovered. There are several rules that apply for proper grouping that extracts feasible components, such as trend and higher frequency oscillations. We suggest to apply the linear operator $\mathcal{H}$ to both sides of the *full* SVD decomposition (7); if $\widetilde{X}_i = \mathcal{H}X_i$ then

$$X = \widetilde{X}_1 + \cdots + \widetilde{X}_d \tag{8}$$

The sum of any two Hankel matrices on the right-hand side of (8) needs not to be Hankel; in this connection, it easily follows from $\langle \widetilde{X}_i, \widetilde{X}_j \rangle_{\mathcal{M}} = 0$ that $\widetilde{X}_i + \widetilde{X}_j$ is Hankel Golyandina et al. (2001) and thus the trajectory matrix of some component time series. By $\langle \cdot, \cdot \rangle_{\mathcal{M}}$ we mean the standard inner product compatible with the

Frobenius-Perron matrix norm. The condition $\langle \widetilde{X}_i, \widetilde{X}_j \rangle_{\mathcal{M}} = 0$ will be referred to as *weak L-separability*. By joining elementary components in (8) having minimum distance in terms of weak L-separability, we often obtain a sensible grouping whose component matrices are as close as possible to Hankel matrices, and, for this reason, feasible of proper interpretation after diagonal averaging. A suitable measure of weak L-separability between components i and j in (8) is the *w-correlation* $w_{ij} = \langle \widetilde{X}_i, \widetilde{X}_j \rangle_{\mathcal{M}} / \|\widetilde{X}_i\|_{\mathcal{M}} \|\widetilde{X}_j\|_{\mathcal{M}}$. If the absolute value of the w-correlation is small, then the corresponding series are almost $\mathcal{M}$-orthogonal, but if the value is large, then the two series are far from being $\mathcal{M}$-orthogonal and thus badly L-separable. Consequently, grouping of components can be made by means of complete link hierarchical clustering assuming $1 - |W| = \left\{1 - |w_{ij}|\right\}_{i,j=1}^{d}$ as the dissimilarity matrix.

An essential target is to estimate the number of clusters p (i.e. the number of RCs). As each eigenvalue λ_i measures the degree of approximation of the component matrix X_i to the trajectory matrix X, we can define an pseudo-R_p^2 score to measure the degree of homogeneity within each cluster. Let $SST = \sum_{j=1}^{p} \sum_{s=1}^{n_j} (\lambda_{sj} - \overline{\lambda})^2$ be the total sum of squares with respect to the full eigenvalue spectrum; similarly, let $SSW_j = \sum_{s=1}^{n_j} (\lambda_{sj} - \overline{\lambda}_j)^2$ be sum of squares within the j-th group (λ_{st} denotes the s-th eigenvalues within the j-th cluster). The R_p^2 index for a decomposition into p groups is defined as

$$R_p^2 = \frac{SST - \sum_{j=1}^{p} SSW_p}{SST} = \frac{SSB_p}{SST} \tag{9}$$

A sensible decision criterion prescribes that a decomposition into $p^\star$ group is chosen if $\sup_p (R_p^2 - R_{p-1}^2)$ is reached for $p = p^\star$, for p varying into a suitable range (for example $p = 3, \ldots, 8$).

3 Results and Discussion

In order to consider Fourier decomposition of the PM_{10} series, cut-off periods $2\pi/\omega_j = N/j$ were respectively set to 579, 30, 14, 7 and 3.5 days; for example, the first component represent the contribution from $\geq$30 days, the second from $14 - < 30$ days and so on. The final result of the decomposition is shown in Fig. 2; estimates of timescale effects are reported in Table 1, from which it is apparent that neither mortality displacement nor associations at longer timescales occur (p-values and model score computations were carried out by the R `mgcv` package, Wood 2004).

In determining an appropriate window length L for SSA decomposition, we computed the smoothed periodogram of the residual PM_{10} series. We found a dominant frequency corresponding to a period of about 26 days. If L is relatively small, component separation results are less stable with respect to small perturbations in L (Golyandina et al. 2001) and thus we tried a range of integer

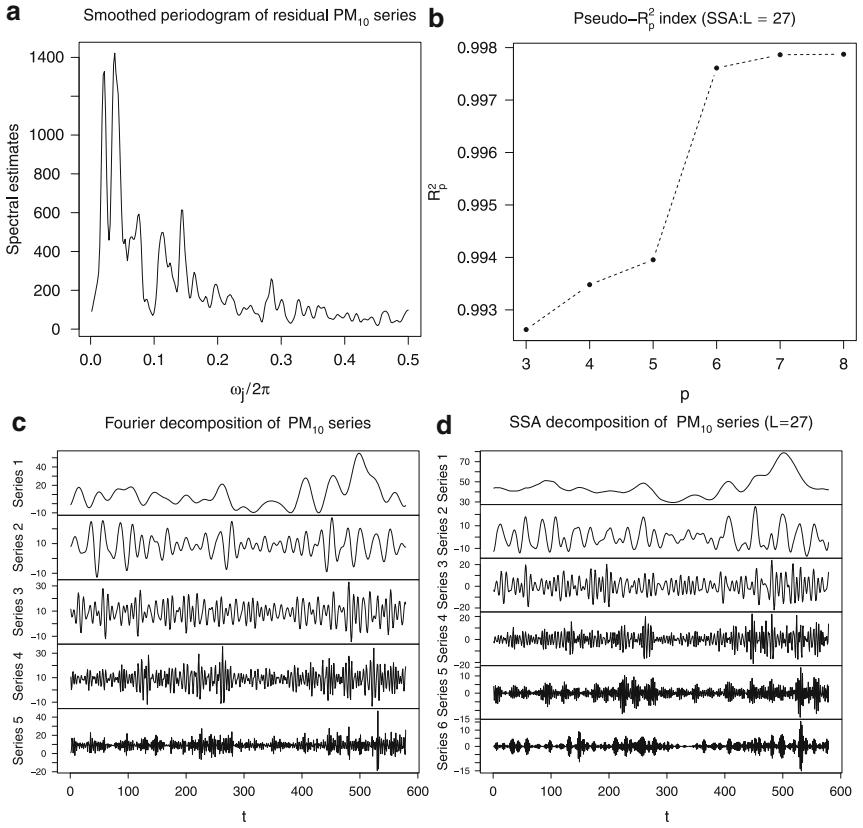


Fig. 2 Output of the decompositions discussed in the text. The periodogram of residual PM_{10} series was smoothed by a series of moving averages of length 3 and 5

values around $L = 26$, comparing them on the ground of the predictive power of the corresponding RCs. The optimal decomposition, in terms of the GCV score, was obtained for $L = 27$. Timescales associated with RCs were estimated as $\Pi_\ell =$ Number of days N/Number of peaks in the ℓ-th RC, for $\ell = 1, \ldots, p$. By exploiting this simple device we found $\Pi_1 = 34.06$, $\Pi_2 = 19.30$, $\Pi_3 = 7.30$, $\Pi_4 = 4.08$, $\Pi_5 = 2.80$ and $\Pi_6 = 2.26$. Estimates of timescale effects are not statistically significant at shorter timescale (see p-values shown in Table 1, but see also the lower GCV score than the first model), but these results suggest a negative association at the longest timescale (about one month). Our results suggest that increases in overall risk associated with airborne pollution are not statistically significant for susceptible populations, and that it is likely to postulate the existence of a pool of healthy individuals which are still healthy one months after exposure.

A weakness of our approach is that the window length L needs to be set in some way; an appropriate choice is crucial, given that results may significantly

Table 1 Timescale effect estimates, p-values and global model scores for the two decompositions discussed in the text

	Fourier decomposition				SSA with $L = 27$			
	Estimate	P-value	ARR	95% C.I.	Estimate	P-value	ARR	95% C.I.
Intercept	2.1972	<2e-16	–	–	2.4779	<2e-16	–	–
Series 1	0.0004	0.790	–	–	−0.0062	0.0494	0.9394	(0.8828, 0.9997)
Series 2	−0.0004	0.827	–	–	0.0027	0.1178	–	–
Series 3	−0.0013	0.467	–	–	−0.0015	0.4315	–	–
Series 4	−0.0012	0.481	–	–	0.0005	0.8216	–	–
Series 5	0.0026	0.186	–	–	0.0020	0.5523	–	–
Series 6	–	–	–	–	0.0017	0.6591	–	–
ϕ	1.3332				1.3247			
GCV	1.3916				1.3868			

vary even though L undergoes small changes. Anyway a satisfactory method for window length setting is still missing. Another possibility that will be explored in future papers is the "Empirical Mode Decomposition", based on the Hilbert-Huang transform (Huang et al. 1999), by means of which any complicated data set can be decomposed into a small-number of intrinsic mode functions and no free parameters need to be set.

Acknowledgements Massimo Bilancia conceived the study, wrote Sect. 2 and revised the draft manuscript. Francesco Campobasso wrote Sects. 1 and 3. Both authors read and approved the final manuscript.

References

Bell, M. L., Samet, J. M., & Dominici, F. (2004). Time-series studies of particulate matter. *Annual Review of Public Health, 25*, 247–280.

Bilancia, M., & Stea, G. (2008). Timescale effect estimation in time-series studies of air pollution and health: A singular spectrum analysis approach. *Electronic Journal of Statistics, 2*, 432–453.

Broomhead, D. S., & King, G. P. (1986). Extracting qualitative dynamics from experimental data. *Physica D, 20*, 217–236.

Buchstaber, V. M. (1994). Time series analysis and Grassmanians. In S. Ginkin (Ed.), *Applied problems of radon transform* (Vol. 162, pp. 1–17). AMS Translations – Series 2. Providence, RI: AMS.

Dominici, F., McDermott, A., Zeger, S. L., & Samet, J. M. (2003). Airborn particulate matter and mortality: Timescale effects in four US cities. *American Journal of Epidemiology, 157*(12), 1055–1065.

Dominici, F., Sheppard, L., & Clyde, M. (2003). Health effects of air pollution: A statistical review. *International Statistical Review, 71*, 243–276.

Golyandina, N., Nekrutin, V., & Zhigliavsky, A. (2001). *Analysis of time series structure: SSA and related techniques*. London: Chapman & Hall/CRC.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N., Tung, C. C., & Li, H. H. (1999). The empirical mode decomposition and the Hilbert Spectrum for non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A, 454*, 903–955.

Samet, J., Zeger, S., Kelsall, J., Xu, J., & Kalkstein L. (1998). Does weather confound or modify the association of particulate air pollution with mortality? *Environmental Research. Series A, 77*, 9–19.

Wood, S. N. (2004). *Generalized additive models: An introduction with R*. Boca Raton, FL: Chapman & Hall/CRC.

Identification of Specific Genomic Regions Responsible for the Invasivity of *Neisseria Meningitidis*

Dunarel Badescu, Abdoulaye Baniré Diallo, and Vladimir Makarenkov

Abstract In this article, we present four distance-based discrimination functions for the identification of relevant genomic segments that distinguish between two groups of data. These discrimination functions are designed for the detection of genomic regions responsible for disease. One of them was previously employed for the analysis of the Human Papilloma Virus family in relation to carcinogenicity (Diallo et al. 2009). Here, we used an improved version of the algorithm described in Badescu et al. (2008) and Diallo et al. (2009) for analyzing the information content of a multiple sequence alignments (MSA) in relation to epidemiologic data. In this study, those functions have been applied to identify specific genomic regions responsible for the hyperinvasivity of *Neisseria Meningitidis*. *Neisseria Meningitidis* is a major causal agent of meningitis and septicaemia worldwide. This study suggests that the tested functions permit to identify relevant regions and known molecular features. We found that one of the new functions tested is specifically well correlated with surface-exposed loops, regions important in vaccine design.

1 Introduction

The evolution of bacteria is driven by several small scale evolutionary events such as substitutions, insertions and deletions and of nucleotides, and large scale mutations such as horizontal gene transfer, duplication of nucleotide segments, etc. However, on a small scale time frame, alleles from the same bacteria organisms diverge little. For instance, when looking into a single gene, only small scale evolutionary changes are commonly present. Alignment of those allele sequences ensure that nucleotides placed on the same region (i.e. same site position), impart the same evolutionary history. Under selective pressure, these molecular modifications will lead

A. Baniré Diallo (✉)
Département d'informatique,Université du Québec à Montréal, C.P. 8888,
Succursale Centre-Ville, Montréal (Québec), H3C 3P8, Canada
e-mail: diallo.abdoulaye@uqam.ca

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_53,

to different epidemiological behaviors. One big issue in comparative genomics is the identification of these molecular modifications through the sequences conservation. One of the well known strategies for identifying genomic sequence regions that have high impact on the given species according to a specific behavior, consists of detecting sequence regions that are conserved across species. Highly conserved regions specific to a family of organisms might have an important role on the common functions of this group (Siepel et al. 2005). Several methods for finding unusual hyper conserved genomic segments have been designed. Most of them are based on phytogenetic trees. They identify hyper conserved genomic segments using hidden Markov model such as Siepel et al. (2005), detect sequences under lineage-specific selection such as DLESS (Siepel et al. 2006) or detect nearly exact motifs using phylogenetic footprinting Blanchette and Tompa (2003). Other simpler methods such as signatures or exact motif finding are also used but they have little application. It is important to notice that, the latter methods analyze a single family at once, and cannot take into account different data categories. Finally due to their exponential time complexity, they are limited to small number of taxa. Being able to classify a family of organisms into a few categories can be an important clue for the detection of their common features. Statistically analyzing the intra- and inter-population variability between two categories can help finding quickly the DNA regions responsible for the difference between the observed categories. In this paper, we tested four distance-based functions for the identification of such differences. They are integrated into an improved version of the algorithm of Badescu et al. (2008) for analyzing the information content of a MSA in relation to epidemiological data. The proposed functions have been applied to the detection of DNA regions related to the hyperinvasivity of the *Neisseria Meningitidis*. The results presented here suggest that the new functions have a good correlation with known molecular features involved in immunological conflict, and responsible for hyperinvasivity.

2 *Neisseria Meningitidis* and the *FrpB* Proteins

Neisseria Meningitidis is a Gram negative bacteria with a high medical importance and very large family. It has small genomic size with 2.2 Mbp. At the time of writing, more than 7,300 genetically distinct known members of Neisseria species were listed into the PubMLST database (Jolley et al. 2004). The latter factor makes it well suited for carrying out for comparative genomic studies. However, bacteria grown under iron starvation express several proteins, *FrpB* being the most abundant one. It is a 70 kDa outer membrane protein (OMP), expressed in large amounts in all strains, and antibodies against this protein appear to be bactericidal. Since iron limitation is a condition met in the body, proteins expressed under this condition are considered as a potential vaccine component (Pettersson et al. 1997). A putative *FrpB* protein topology was proposed (Pettersson et al. 1995) with a 26-stranded β-barrel and a 22-stranded β-barrel with 11 surface-exposed loops. It is these loops that are accessible to the host immune system. Natural antibodies are generated

against these regions and bacteria express variability in order to evade this defence mechanism. Also these 11 surface-exposed loops are also a favorite place of guest-host interaction. This study will focus on the detection of these surface-exposed loop regions under the knowledge of the organism categories (invasive and non-invasive alleles).

3 Algorithm for Detection of Genomic Regions Responsible for Disease

This section describes the steps used for finding genomic regions that responsible for the invasivity of *Neisseria Meningitidis*. The algorithm tests several hypothesis such as whether sequence regions responsible for invasivity are likely to be more similar among invasive strand, or not. The algorithm takes as input a multiple sequence alignment (MSA) of nucleotides, and a set of organisms, clustered into two different groups (i.e. categories) according to their invasivity: X (invasive) and Y (non-invasive). We scanned the sequence alignment using an overlapping sliding window of a fixed width (in our experiments the window width ranged from 5 to 20 nucleotides). Once the window position in MSA is fixed and the organisms are assigned to the groups X and Y, various discrimination functions can be defined. The different steps of our procedure are described below. Figure 1 presents the algorithmic flow of the hit identification, followed by the description of the different steps of this algorithm.

Step 1: Collection and annotation of the MSA of the FetA alleles: MSA of the *FetA* allele sequences are available from the Neisseria Research Community databank (Thompson et al. 2003; Neisseria Research Community Website 2009). We annotated the MSA using the information on the surface-exposed loops, beta-sheets and periplasmic loops, as was explained in Kortekaas et al. (2007). Identification and presentation were carried out on the H44/76 strain, with the GenBank accession number X89755.1 (Pettersson et al. 1995).

Step 2: Classification of taxa as invasive or non-invasive: To form groups X and Y on an invasivity basis we used a list of identified hyperinvasive meningococci (Urwin et al. 2004). We built a list of uniques *FetA* sequence tags carried by these alleles. With a local BLAST we searched for the presence of those tags into the distinct sequences belonging to the MSA (Altschul et al. 1990). We classified as belonging to the X category any allele that has a perfect hit with at least one of the selected invasive tags. All others were put in the non-invasive category Y.

Step 3: Computating the detection functions Q values: For a fixed alignment window position the hit region identiPcation functions (i.e. hit region is a region responsible for disease), denoted here as Q_1, Q_2, Q_3 and Q_4 are computed as follows. These functions are defined as a difference between the means of the squared distances computed among the sequence fragments (bounded by the sliding

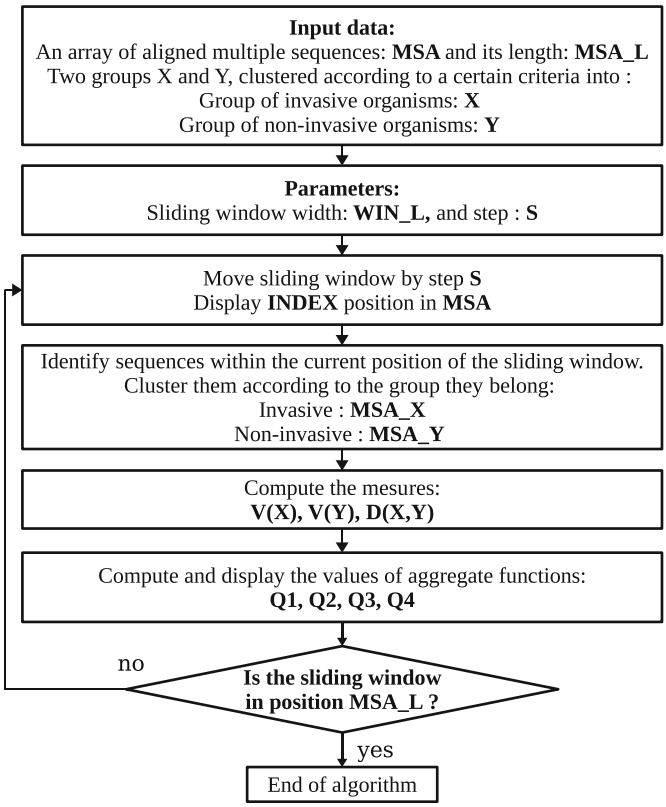


Fig. 1 Algorithmic flow of the hit identification function Q, using pluggable functions Q_1, Q_2, Q_3, Q_4

window position) of the taxa from the set X and those computed only between the sequence fragments from the distinct sets X and Y. To compute the function values, the variability of groups X and Y respectively $V(X)$ and $V(Y)$ is computed as well as the distance between X and Y, denoted $(D(X, Y))$. The variability of the group X corresponds to the mean squared distance computed among the sequence fragments of the invasive organisms. This variability is computed as follows:

$$V(X) = \frac{1}{(N(X) \cdot (N(X) - 1)/2)} \sum_{\{x_1, x_2 \in X | x_1 \neq x_2\}} dist_h^2(x_1, x_2). \quad (1)$$

The variability of class Y corresponds to the mean of the squared distance computed among the sequence fragments of the non-invasive organisms. This variability is computed as follows:

$$V(Y) = \frac{1}{(N(Y) \cdot (N(Y) - 1)/2)} \sum_{\{y_1, y_2 \in Y | y_1 \neq y_2\}} dist_h^2(y_1, y_2). \quad (2)$$

The distance between the groups X and Y corresponds to the mean of the squared distances computed among the sequence fragments from X and Y. This distance is computed as follows:

$$D(X, Y) = \frac{1}{N(X) \cdot N(Y)} \sum_{\{x \in X, y \in Y\}} dist_h^2(x, y), \quad (3)$$

where $N(X)$, $N(Y)$, $dist_h^2(x, y)$ are respectively the cardinalities of the groups X and Y, and the square of the Hamming distance between the sequences x and y.

We propose to examine four different hit identification functions allowing one to detect DNA zones responsible for disease. The function Q_1 focuses on the specific regions of the alignment that are either well-conserved within the invasive set X, when it is positive, or highly divergent, when it is negative. The function Q_2 focuses on the specific regions of the alignment that are either well-conserved within the non-invasive set Y, when it is positive, or highly divergent, when it is negative. The function Q_3 is positive when the distance between the taxa in X and Y is higher than the variability between the groups. The last function Q_4 consists only of the mean squared distances between the two groups:

$$Q_1 = D(X, Y) - V(X), \quad (4)$$

$$Q_2 = D(X, Y) - V(Y), \quad (5)$$

$$Q_3 = 2D(X, Y) - V(X) - V(Y), \quad (6)$$

$$Q_4 = D(X, Y). \quad (7)$$

Step 4: Identify hit regions. To identify a region as a hit, one might use a measure to determine whether the given region has a value of Q higher than a predefined threshold. However, it is necessary to normalize the obtained results given by Q_1, Q_2, Q_3 and Q_4 prior to compare them. We compare the trends of the different function according to the known regions of surface-exposed loops of *FetA* alleles. One can also determine the hit regions computing the p-values of the proposed functions (Diallo et al. 2009).

4 Results and Discussion

We scanned the MSA of the *FrpB* gene using the algorithm described in the previous section, with the four versions of the aggregate discrimination function Q, and window of size 10 nucleotides. Larger window sizes were less discriminative in terms of

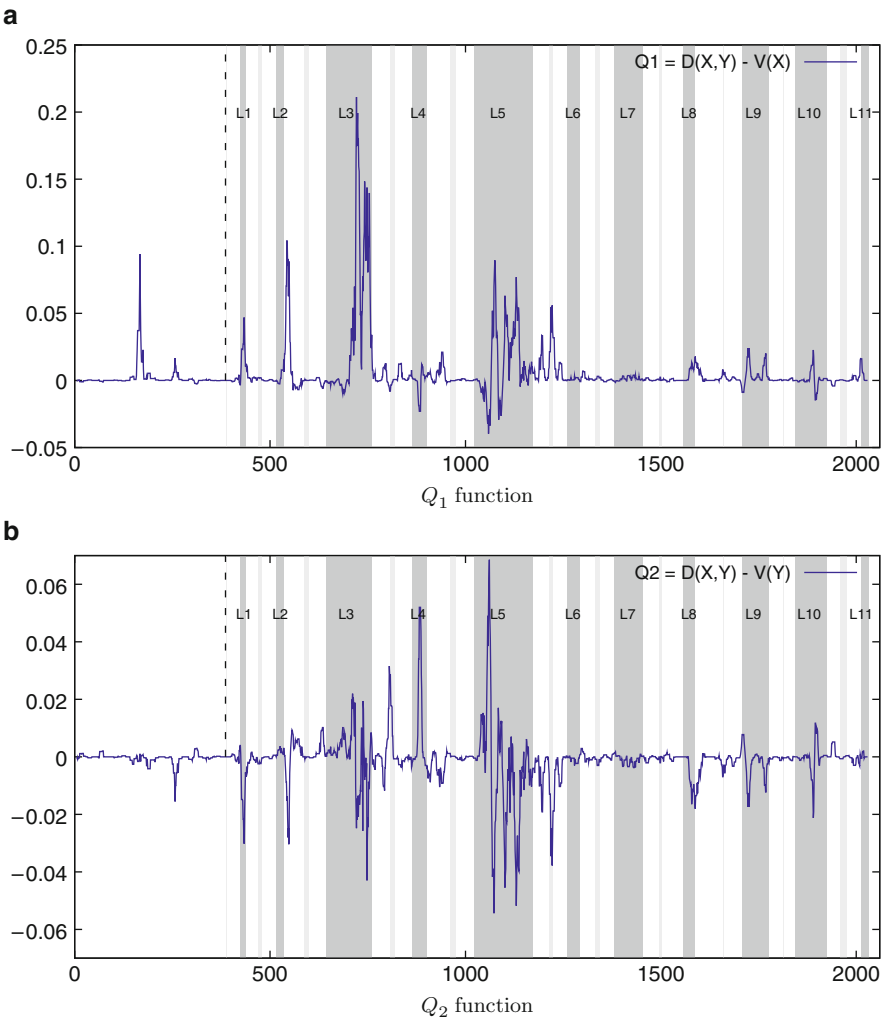


Fig. 2 The variation of the hit identification functions Q_1 and Q_2 for the *Neisseria Meningitidis* containing invasive sequence tags obtained with a non-overlapping sliding window of size 10 during the gene *FrpB* scan. The abscissa axis represents the window position. Gray zones are the positions of the surface-exposed loops

regionality, smaller sizes introduced more noise. One can notice (see Figs. 2 and 3) that the high values of the four tested functions usually correspond to the gray zones (regions supposed to be responsible for the invasivity) of the graphics. The values of the function Q_2 are generally lower than those of the other functions. The latter means that the divergence in Y is almost similar to the divergence between the alleles from X and Y (Fig. 2(b)). The high values of the function Q_1 (above 0.05) in Fig. 2(a) can be induced by highly conserved features in the X. The function Q_3

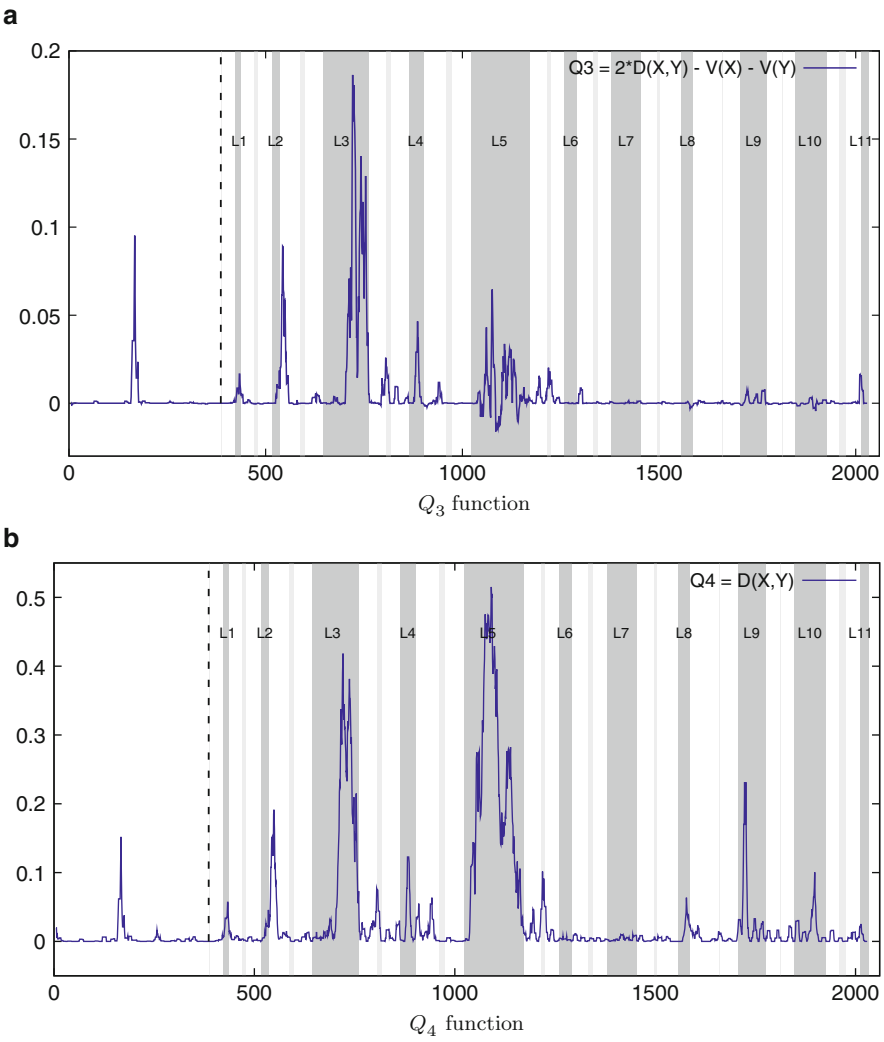


Fig. 3 The variation of the hit identification functions Q_3 and Q_4 for the *Neisseria Meningitidis* containing invasive sequence tags obtained with a non-overlapping sliding window of size 10 during the gene *FrpB* scan. The abscissa axis represents the window position. Gray zones are the positions of the surface-exposed loops

(Fig. 3(a)) correlates less with the gray zone, except for L_3. Furthermore, the trends of the function Q_4 suggests that for almost all gray zones (except L_7), the genomic segments are different between the groups X and Y. In order to compare the values of the four competing functions, we carried out a zero mean and unit variance normalization. Table 1 presents the maximum values obtained for to the eleven gray zones in Figs. 2 and 3. More than the half of the maximum values on this table are

Table 1 Normalized maximum values of the functions Q_1, Q_2, Q_3, Q_4 in each gray region. Higher values for each region are highlighted

MAX	1	2	3	4	5	6	7	8	9	10	11	Gray zones detected
Q_1	4.51	6.92	21.34	0.98	8.9	0.07	0.15	1.57	2.19	2.06	0.24	6
Q_2	−0.57	0.9	4.56	10.56	13.87	0.74	0.34	0.17	1.71	2.5	0.57	4
Q_3	2.12	8.56	26.46	6.42	9	0.52	−0.05	−0.12	0.84	0.2	0.75	5
Q_4	2.51	5.19	21.56	5.99	26.66	0	0.05	2.87	11.72	4.82	0.05	8

Last column shows the number of detected regions scoring over the threshold 2.0

above the fixed threshold of 2. This result shows that the function Q_4 is the best one in terms of extracellular loops (gray zone) detection. The same conclusion can be drawn when observing the graphics in Figs. 2 and 3.

5 Conclusion

In this paper we considered four different functions for detecting hit regions responsible for disease. We found that the function Q_4 correlate the best with the surface exposed loops (a feature of the secondary structure of OMPs) of the *Neisseria Meningitidis* of the *FrpB* gene. This suggests that our algorithm is able to detect known regions of interest in respect to given epidemiological criteria. Another interesting development would be to design a statistical test allowing one to measure the signiÞcance of the obtained results such as computing p-values. It will be also important to test the four functions considered in this study, in the other context where the information about the species can be grouped into categories according to specific features such as biological functions, phenotypic differences or behavioral changes.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 403–410.

Badescu, D., Diallo, A. B., Blanchette, M., & Makarenkov, V. (2008). An evolution study of the human papillomavirus genomes. In *Proceedings of RECOMB Comparative Genomics 2008, Springer, Lecture Notes in Bioinformatics Series, Paris*, 128–140.

Blanchette, M., & Tompa, M. (2003). FootPrinter: A program designed for phylogenetic footprinting *Nucleic Acids Research, 31*(13), 3840–3842.

Diallo, A. B., Badescu, D., Makarenkov, V., & Blanchette, M. (2009). A whole genome study and identification of specific carcinogenic regions of the Human Papilloma Viruses. *Journal of Computational Biology, 16*(10): 1461–1473.

Jolley, K., Chan, M.-S., & Maiden, M. (2004). mlstdbnet – distributed multi-locus sequence typing (mlst) databases. *BMC Bioinformatics, 5*(1), 86.

Kortekaas, J., Pettersson, A., van der Biezen, J., Weynants, V. E., van der Ley, P., Poolman, J., Bos, M. P., & Tommassen, J. (2007). Shielding of immunogenic domains in Neisseria Men. *FrpB* by the major variable region. *Vaccine, 25*(1), 72–84.

Neisseria Research Community Website. (2008). From http://www.neisseria.org.

Pettersson, A., Poolman, J. T., van der Ley, P., & Tommassen, J. (1997). Response of Neisseria Men. to iron limitation. *Antonie van Leeuwenhoek, 71*, 129–136.

Pettersson, A., Maas, A., van Wassenaar, D., van der Ley, P., & Tommassen, J. (1995). Molecular characterization of *FrpB*, the 70-kilodalton iron-regulated outer membrane protein of *Neisseria Meningitidis*. *Infection and Immunity, 63*(10), 4181–4184.

Siepel, A., Pollard, K. S., & Haussler, D. (2006). New methods for detecting lineage-specic selection. *Proceedings of the 10th International Conference on Research in Computational Molecular Biology (RECOMB 2006)* (pp. 190–205).

Siepel, A., Bejerano, G., Pedersen, J. S., et al. (2005). Evolutionarily cons. elements in vertebrate, insect, worm, and yeast genomes. *Genome Research, 15*, 1034–1050.

Thompson, E. A. L., Feavers, I. M., & Maiden, M. C. J. (2003). Antigenic diversity of meningococcal enterobactin receptor *FetA*, a vaccine component. *Microbiology, 149*(7), 1849–1858.

Urwin, R., Russell, J. E., Thompson, E. A. L., et al. (2004). Distribution of surface protein variants among hyperinvasive meningococci: Implications for vaccine design. *Infection and Immunity, 72*(10), 5955–5962.

Classification of ABC Transporters Using Community Detection

Claire Gaugain, Roland Barriot, Gwennaele Fichant, and Yves Quentin

Abstract The ATP-binding cassette (ABC) transporters are one of the major family of active transporters, present in the three domains of life. They are involved in the uptake and efflux of a wide variety of compounds, and thus play an essential role in the adaptation of organisms to their environment. We propose a new approach of ABC system classification exploiting the large amount of data available through systematic genome sequencing. Our main goal is to refine their functional prediction, i.e., which compound(s) is transported by a system. Our method relies on the identification of orthologous genes, i.e., genes that have diverged from a common ancestor after an event of speciation and have retained the same function. These evolutionary links are inferred from sequence comparisons. A graph is constructed where vertices represent genes and edges orthology links. Our initial hypothesis is that highly connected isolated groups in our graph correspond to ABC genes involved in the transport of the same compound(s). However, we obtained much more complex graphs due to complex evolution scenarios. Thus, we have tested community detection algorithms to identify highly connected sets of orthologous genes in our graphs. The approach presented here enables us to study entire families of ABC transporters and to bring out communities that correlate well with subdivisions previously identified with another method. The communities obtained are then described in terms of function, i.e., specific compound(s) transported, using the functional annotations available in TCdb.

1 Introduction

ATP-Binding Cassette (ABC) systems represent one of the largest family of proteins widespread in the three kingdoms of life (for a recent review, see Davidson et al. 2008). They are characterized by the presence of at least one ATP-hydrolysing

Y. Quentin (✉)
Université de Toulouse, UPS, Laboratoire de Microbiologie et Génétique Moléculaire, F31000 Toulouse, France
e-mail: Yves.Quentin@ibcg.biotoul.fr

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_54,

domain (or Nucleotide Binding Domain, NBD). They can be divided in three groups, according to their function: exporters, importers and systems involved in processes not related to transport. Exporters are found in prokaryotes and eukaryotes. In humans, they contribute to various human diseases and multidrug resistance. In prokaryotes, they are implicated in the export of cell waste products, toxins, cell surface components and molecule involved in bacterial pathogenesis. Importers are present only in bacteria and archaea, and participate in the uptake of nutrients necessary for metabolism and cell integrity. Despite the heterogeneity of molecules carried, ABC transporters share a common architecture composed of four functional domains: two Nucleotide Binding Domains (NBDs) that bind and hydrolyse ATP to supply energy for the transport, two Membrane Spanning Domains (MSDs) that form the pore for the transit of molecules and confer specificity. The importers require a supplementary protein, a Solute-Binding Protein (SBP), that captures specific substrate(s) with high affinity and contribute to the regulation of the NBD activity. It has been shown that the specificity of substrate is closely related to the sequence conservation of the ABC partners and different specific classification schemes based on sequence similarity have been proposed (Davidson et al. 2008; Fichant et al. 2006; Saier 2000). They agree on a set of about 30 subfamilies that diverged very early in the evolution. In this paper, we address the question of ABC system classification throughout an original approach based on the identification of communities (vertices that are more densely interconnected) into networks (Newman and Girvan 2004). The results obtained reveals a community structure into subfamilies previously described. This subfamily refinement can be correlated to new or more specific substrates when experimental data are available.

2 Materials and Methods

2.1 Data Sources

The completely sequenced genomes of 150 prokaryotes used in this work were downloaded from the EBI (http://www.ebi.ac.uk/genomes/). The protein sequences of the ABC transporters were retrieved from a local version of ABCdb (http://www-abcdb.biotoul.fr). This database is manually curated, and include information on ABC transporter partners - subfamily prediction, domain organization – and the assembly of partners in ABC systems (Fichant et al. 2006; Quentin et al. 2002). They are organized in 21 main subfamilies: 10 importers, 7 exporters and 4 “non-transporters”. This represents about 12.500 proteins and hundred thousands isorthologous links.

2.2 Methods

Our approach is based on the concept of homology. There are two types of homology: paralogy and orthology. Paralogous genes are genes that have diverged from a common ancestor after a duplication (two copies in a genome). Orthologous genes are genes that have diverged from a common ancestor after a speciation event (evolutionary process by which new species arise), and it is assumed that these genes have retained the same function. Since we do not have direct access to these evolutionary relations, they must be inferred from sequence comparisons.

2.2.1 Identification and Filtering of Isorthologous Links

A minimal definition of orthology is the bi-directional best hit (BBH) criterion, where two genes *a* and *b* from genomes *A* and *B*, are considered to be orthologs if *a* is the best hit of *b* in genome *A* and reciprocally. This definition is more restrictive than the one proposed by Fitch (1970), since each gene can have only one ortholog in another genome, but it is also less accurate since we cannot exclude that a gene actually has more than one ortholog in another genome. Fitch (2000) proposed isorthologs to designate orthologous pairs of genes that are unequivocally related by an orthologous relationship. Thus, isorthologous genes have higher chances to have retained the same function in both genomes and groups of isorthologs would define sets of functionally related proteins. Therefore, we added supplementary constraints on BBH criterion to identify isorthologs: if *a* or *b* has a paralogous gene named *c* then, if the score of *a* versus *b* is higher than the score of *a* or *b* versus *c* then the *a* and *b* are isorthologs, otherwise *a* and *b* are considered as orthologs (Overbeek et al. 1999).

However, the high level of paralogy in ABC family suggests that several rounds of duplication, deletion and/or lateral gene transfer events occurred along the evolution, probably in response to variations of environmental conditions. This instability during evolution and/or intrinsic errors in the method to predict homology links lead to errors in the detection of isorthologous relationships. In order to reduce such errors in our data, we imposed that each protein belongs to at least one triangle ("transitive" relationships) in order to remove recent duplications and misleading links.

2.2.2 Identification of Isortholog Groups by Community Detection

The isorthology and orthology relationships inferred between gene/protein pairs are not transitive. Therefore, there is not a straightforward approach to compute groups of orthologous genes. The COG database was the first attempt to classify proteins from completely sequenced genomes on the basis of the orthology relationships (Tatusov et al. 1997). The overall COG construction process relies on a clustering method based on the agglomeration of mutually consistent triangles of orthologous

relationships. Additional procedures are used such as the addition of species-specific paralogs, as well as manual modifications like the recognition and separation of protein domains and refinement of big clusters in more compact ones, especially for large protein families like ABC transporters. The manual refinement of the complex clusters is time consuming and not based on the intrinsic properties of the data. So we decided to experiment network partitioning methods on isorthology networks. The approach called community detection retained our attention, since it was widely used these last years to find highly connected groups (communities) in biological and social networks (Girvan and Newman 2002). Community detection methods enable the identification of vertices that are more densely interconnected among each other than with the rest of the network. In order to assess the relevance of this approach, we used the community detection algorithms implemented in the R (http://www.R-project.org) in the package "igraph" (Csardi and Nepusz 2006; http://igraph.sourceforge.net/). Several algorithms are available and they rely on the optimization of the modularity Q which quantifies the quality of a network partition (Newman and Girvan 2004). Two agglomerative methods, Clauset et al. (2004) and Pons and Latapy (2006) have been tested:

- Fastgreedy algorithm begins with each vertex in a different community (we have as many communities as vertices); it repeatedly merged the two communities whose join leads to the largest increase of modularity Q. These successive merges can be represented as a dendrogram: leaves are the initial step (one vertex=one community), and the internal nodes are the merges. It produces a hierarchical decomposition of the network at different levels. This algorithm exploits some shortcuts to find the best modularity which is costly in time, and thus provides us a fast implementation of community detection, usable for large networks.
- Walktrap algorithm is based on the principle of random walks that tend to keep trapped in dense subnetworks (communities). Short random walks (five steps) are performed on the network to determine the transition probability from a vertex i to a vertex j through 5 steps. From this probability we can obtain a distance between each pair of vertices. The algorithm starts from a partition where each vertex is a community. Two communities are merged such as distances within communities are smaller than those between communities. Distances between communities are updated and these two steps are performed until one community containing all vertices is obtained. At each iteration, we have different partitions of the network into communities and Q is calculated for each of them. The partition that gives the highest value of Q is kept.

2.2.3 Validation

During evolution, partners of an assembly are assumed to co-evolve in order to maintain their interactions and keep fulfilling their function. Hence, we expect that the partitions – communities of isorthologous proteins – obtained independently on each type of partners will be correlated. In other words, a community of isorthologous proteins, such as NBDs for example, should be involved in a set

of "isorthologous systems". In turn, these systems involve the other partners, i.e. MSDs and SBPs, which are expected to cluster together in their respective communities of MSDs or SBPs. This property provides an external criterion to validate the partitioning procedure results.

For this, we use a set of curated assemblies of ABC system partners (Fichant et al. 2006; Quentin et al. 2002) and we check the frequency of how often corresponding communities of MSDs, NBDs and SBPs include all the partners of the curated assemblies. The community detection algorithm yielding the highest frequency of complete assemblies re-formed is considered the most relevant. To visualize the correlation between identified communities and curated assemblies, we use a binary matrix (communities x assemblies) in which a cell contains 1 if a protein of the community (row) is involved in the assembly (column), and 0 otherwise. Then, a clustering is applied on communities (rows). If a good correlation exists between communities and assemblies, then we should observe clusters corresponding to classes of isorthologous ABC systems (see Fig.1 for an illustration). As a shortcut, we denote this visualization a heatmap in the rest of this manuscript, though the color of the cells only reflects the presence/absence of proteins in a community and in an assembly. Moreover, in order to strengthen the communities found, we retrieve functional annotations of ABC transporters available in TCdb (Saier et al. 2006), and we search for conserved genes in the neighborhood of the genes encoding ABC partners.

Fig. 1 Network partitions and heatmap obtained for subfamily of pentose-related importers. On the *left*, in each network (NBDs, MSDs, SBPs), *grey* areas surrounding network vertices correspond to communities identified by Walktrap. On the heatmap (*right*), each column corresponds to an assembly, and each line to a community. Taxonomy (*black* for archaea and *grey* for bacteria) is reported on a line above the heatmap. Predicted substrates are indicated on the *right*

3 Results

3.1 General Results on ABC System Classification

Despite the different network structures obtained on each subfamily and on each ABC partner, both algorithms (Fastgreedy and Walktrap) were able to detect very similar communities. This confirms the existence of such communities, and at the biological level, the organization of ABC subfamilies in more specific classes. Unfortunately, the scarcity of experimental evidences leaves many classes without substrate prediction. Nevertheless, a general trend seems to emerge from our results: among many subfamilies only one class covers both bacteria and archaea proteins and could correspond to the ancestral systems.

3.2 Results on Pentose-Related Importer Subfamily

We choose to present the results obtained on the subfamily known to import pentose-related subtrates. After filtering the isorthologous links, three networks are obtained, respectively composed of 87 NBDs connected by 1084 isorthologous links, 160 MSDs connected by 2,305 isorthologous links, and 99 SBPs connected by 1,118 isorthologous links. Both partitioning algorithms (Walktrap and Fastgreedy) have been applied on these networks. The same partition into 9 and 14 communities have been obtained by the two algorithms on the NBD and MSD networks, respectively. Whereas two different partitions are found in the SBP network (8 and 10 communities for Fastgreedy and Walktrap respectively). The results are summarized in Fig. 1. The heatmap highlights clusters of communities that are correlated to classes of curated ABC assemblies. The Walktrap and Fastgreedy algorithms recover 80% of the assemblies. Altogether, these results strengthen the confidence in the partitions obtained.

Six classes have been identified. Functional annotations retrieved from TCdb allows substrate prediction for four of them: autoinducer-2 (AI2) for systems of cluster 1, xylose for systems of cluster 2, galactose/glucose for systems of cluster 5, and ribose for systems of cluster 6.

A conserved gene coding for ribokinase has been identified in the vicinity of ABC protein-encoding genes of cluster 6, hence reinforcing our prediction (see Fig. 2).

For each system of cluster 2, computation of phylogenetic profiles (Enault et al. 2005) reveals, for more than half of the genomes, the co-occurrence of two enzymes (high mutual information score) whose functions are xylose isomerase and xylulokinase. Again, this analysis strengthens the prediction that xylose is imported by these transporters.

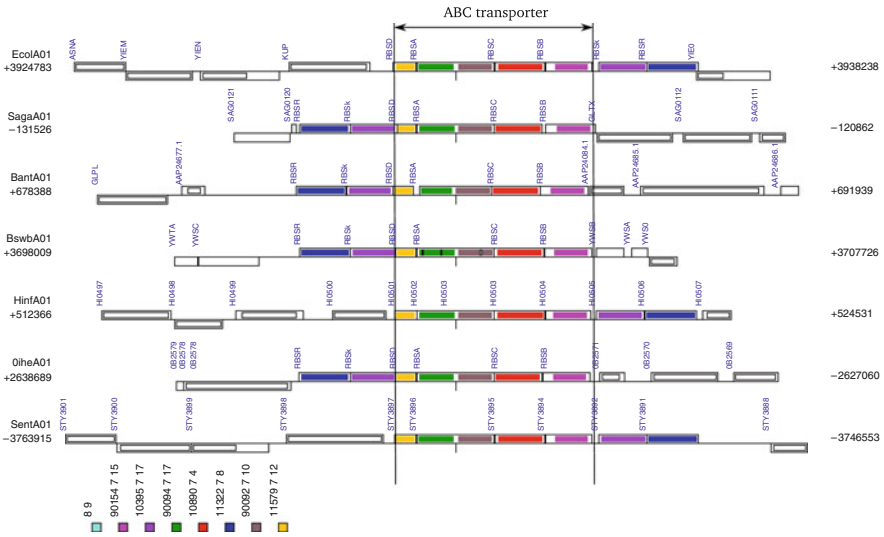


Fig. 2 Neighbourhood of the ribose ABC transporters. Each line is a part of a genome (name of the organism on the *left*). This representation shows genes that are neighbours on the chromosome, and the grey shades correspond to different functional annotation. From left to right and between the two vertical lines, genes encoding SBP, NBD and MSD. The two darkened neighbours (on the *left* or *right* of the vertical lines) are involved in the metabolism of ribose (ribokinase and ribose operon repressor)

The functional prediction that cluster 3 systems may be involved in ribonucleoside transport is weaker than the previous ones as it is based only on sequence similarities with such transporters of TCdb.

From an evolutionary point of view, the taxonomic distribution of the different classes reveals that only cluster 3 is present both in bacteria and archaea. Therefore, we can hypothesize that this cluster corresponds to the ancestral system and that the others appeared more recently in bacteria.

4 Conclusion

In this paper, we present a generic strategy to aid the functional classification of proteins belonging to large multigenic families involved in integrated systems. Its original principle is to identify correlated clusters of proteins among networks of isorthologous proteins via a community detection approach. Very promising results were obtained by directly applying the strategy to ABC transporters despite the complexity of this protein family. Indeed, we show that previous classification schemes manually defined by experts could be refined: specific compound(s) can be associated to each new class of transporters, instead of a type of substrate. The ability of the strategy to tackle the high level of paralogy of large protein families opens new

perspectives of research. For ABC transporters, new finer studies will improve our understanding of prokaryote adaptation to their environment, and will also benefit human health as they are involved in pathologies. Our strategy will also be beneficial to the study of other integrated systems such as two component systems. Other community detection algorithms and partitioning methods will be used to compare to these first results and confirm that the approach applied here is the most relevant.

References

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E., 70*(6), 066111.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695.

Davidson, A. L., Dassa, E., Orelle, C., & Chen, J. (2008). Structure, function, and evolution of bacterial ATP binding cassette systems. *Microbiology and Molecular Biology Reviews, 72*(2), 317–364.

Enault, F., Suhre, K., & Claverie, J.-M. (2005). Phydbac gene function predictor: A gene annotation tool based on genomic context analysis. *BMC Bioinformatics, 6*(1), 247.

Fichant, G., Basse, M.-J., & Quentin, Y. (2006). ABCdb: An online resource for ABC transporter repertories from sequenced archaeal and bacterial genomes. *FEMS Microbiology Letters, 256*(7), 333–339.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systems Zoology, 19*(2), 99–113.

Fitch, W. M. (2000). Homology: A personal view in some of the problems. *Trends in Genetics, 16*(5), 227–231.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences USA, 99*(12), 7821–7826.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E., 69*(2), 026113.

Overbeek, R., Fonstein, M., Souza, M. D., Pusch, G. D., & Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences USA, 96*(6), 2896–2901.

Pons, P., & Latapy, M. (2006). Computing communities in large networks using random walks. *Journal of Graph Algorithms and Application, 10*(2), 191–218.

Quentin, Y., Chabalier, J., & Fichant, G. (2002). Strategies for the identification, the assembly and the classification of integrated biological systems in completely sequenced genomes. *Computers and Chemistry, 26*(5), 447–457.

Saier, M. H. (2000). A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiology and Molecular Biology Reviews, 64*(2), 354–411.

Saier, M. H., Tran, C. V., & Barabote, R. D. (2006). TCDB: The transporter classification database for membrane transport protein analyses and information. *Nucleic Acid Research, 34*(Database issue), 181–186.

Tatusov, R. L., Koonin, E. V., & Lipman, D. J. (1997). A genomic perpective on protein families. *Science, 278*(5338), 631–637.

Estimation of the Number of Sustained Viral Responders by Interferon Therapy Using Random Numbers with a Logistic Model

Shinobu Tatsunami, Takahiko Ueno, Rie Kuwabara, Junichi Mimaya, Akira Shirahata, and Masashi Taki

Abstract Infection with hepatitis C virus (HCV) is frequently observed among Japanese hemophilic patients, and critical hepatic diseases arising from HCV infection have become a major cause of death. Therefore, we tried to estimate the benefits of the interferon therapy among 1,241 hemophilic patients with chronic hepatitis. We used the viral genotype and the RNA concentration as two predicting variables for the efficacy of the interferon therapy, and assumed a binomial logistic regression model. The missing data of the patient's viral genotype and RNA concentration were substituted by random numbers simulating the actual observed distribution. By repeating the computation 1,000 times using different sets of random numbers, we estimated the number of sustained viral responders (SVR) resulting from the therapy. We observed certain changes in the estimated number of SVR by changing the dependence assumption of therapeutic efficacy on the predicting variables. In the most optimistic scenario, the estimated number for SVR was 692 ± 17 ($55.8 \pm 1.4\%$), while it was 461 ± 16 ($37.1 \pm 1.3\%$) in case of the most pessimistic scenario. The effect of the missing data on the estimates was not large. Therefore, these estimates will be helpful for making a prospective evaluation of the survival benefits coming from the spread of the interferon therapy.

1 Introduction

A considerable fraction of Japanese hemophiliacs who were born before the establishment of preventive technology for hepatitis C virus (HCV) contamination in the production of coagulation concentrates have been infected with HCV (Taki et al. 2003). As a result, critical hepatic diseases such as cirrhosis, liver failure and hepatocellular carcinoma arising from HCV infection, have become a major cause of

S. Tatsunami (✉)
Unit of Medical Statistics and Institute of Radioistope Research, St. Marianna University School of Medicine (Collaborated with the Research Committee for the National Surveillance on Coagulation Disorders in Japan), Kawasaki 216-8511, Japan
e-mail: s2tatsy@marianna-u.ac.jp

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_55,

death as reported in the annual report on hemophiliacs (Japanese Foundation for AIDS Prevention 2008). Therefore, we have been suggesting that hemophiliacs infected with HCV receive therapy using interferon before the development of hepatic disease with the elapse of time. However, the benefit from the therapy among the population of hemophiliacs has never been estimated quantitatively. Therefore, we tried to estimate the number of patients that will obtain successful results among the Japanese hemophiliac patients.

2 Subjects and Model Assumptions

Regarding the present subjects for the computation, we chose Japanese patients with blood coagulation disorders, mainly hemophiliacs infected with HCV. The patients suffering from critical liver diseases were not included because the therapy with interferon is thought applicable mainly for patients without any critical liver disease. Patients with HIV infection were also excluded because the interaction between HIV and HCV has been known generally (Daar et al. 2001).

Therefore, we included only a total of 1,241 HCV-infected patients that were in any states of chronic hepatitis.

We assumed that the probability of the therapy success crucially depended on the HCV subtype and its RNA concentration in each patient, and the logistic function (Hosmer and Lemeshow 2000) including these two factors as independent variables reliably predicted the probability of the therapy success. The success of therapy was defined as the attainment of sustained viral response with the disappearance of viral RNA in a period longer than 6 months and recovery of liver function. In this context, the benefit of the interferon therapy was evaluated by the number of sustained viral responders (SVR).

3 Methods

Under the assumptions above, we computed the probability of therapy success by the logistic function $f(x_1, x_2)$ as follows:

$$f(x_1, x_2) = \frac{1}{1 + e^{-(B_0 + B_1 x_1 + B_2 x_2)}} \tag{1}$$

where x_1 and x_2 are the viral subtype and RNA concentration of a patient, respectively.

Then we generated a uniform random number, r_i, for the i-th patient, and expressed the success or non-success of the therapy n_i by the following relation:

$$n_i = \begin{cases} 1 & r_i \leq f(x_1, x_2), \\ 0 & r_i > f(x_1, x_2) \end{cases} \tag{2}$$

Finally we computed the number of SVR with the therapy N as follows;

$$N = \sum_{i=1}^{1241} n_i \tag{3}$$

We classified the subtypes of HCV into two types: type 1 and others. Type 1 included viral genotype 1a, and 1b; others 2a, 2b, 3a and 4a. When 1a or 1b was found in a patient with plural viral genotypes, the viral type was defined as type 1; otherwise it was classified as others. Therefore the predicting variable x_1 was expressed by the dichotomous variable 1/0. The unit for another predicting variable x_2 was kilocopies/mL in the original data from the surveillance; however, we substituted its logarithm into (1) after dividing the original value by 10. Further, we substituted 1.0 to x_2 when the RNA concentration was lower than 100 kilocopies/mL.

The missing data for the HCV subtype or RNA concentration in a patient were substituted by random numbers that simulated the observed classification of viral subtype and RNA concentration among the patients with reported values.

By repeating the computation of (3) using different sets of random numbers 1,000 times, we obtained the mean standard deviation, and median of N with respect to six sets of different values of B_0, B_1 and B_2.

4 Results

Complete observations were available in 490 (39.5%) of the 1,241 patients. Viral genotype information was incomplete in 189 (15.2%), RNA concentration in 67 (5.4%) and both were unavailable in 495 (39.9%). The observed frequency of the HCV subtypes is illustrated in Fig. 1. In short, the infection with type 1 virus including mixture types was observed in 415 patients, while the remaining 142 patients with other types. Therefore, we used random numbers with a distribution frequency of 3:1 for the virtual viral type 1 and others; then we substituted them for patients without any information of viral subtypes.

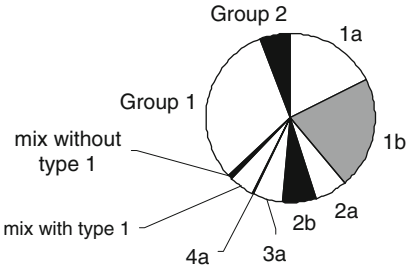


Fig. 1 Distribution of HCV subtypes among the present subjects. Group 1 and 2 means cerotype 1 and 2, respectively, reported for patients without data of viral genotype

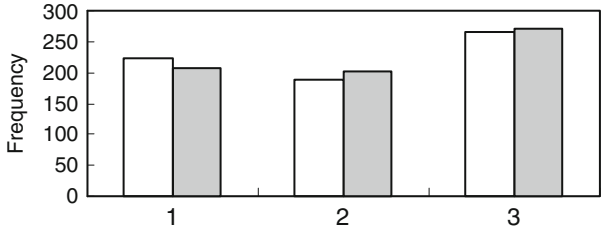


Fig. 2 Frequency distribution of HCV RNA concentrations observed among 679 patients (*blank column*) and 10,000 generated random numbers (*gray column*). Positions 1, 2 and 3 are viral RNA concentrations lower than 100 kilocopies/mL including results under the detectable limit (1), from 100 to 1,000 kilocopies/mL (2), and higher than 1,000 kilocopies/mL (3), respectively. The scale for the random numbers is adjusted for comparison

Table 1 Values of parameters B_0, B_1 and B_2 used in the computation

Parameters	Scenario No.					
	1	2	3	4	5	6
B_0	2.9	3.4	3.9	2.9	2.9	2.9
B_1	−0.9	−0.9	−0.9	−0.7	−0.9	−0.7
B_2	−1.7	−1.7	−1.7	−1.7	−1.5	−1.5

Table 2 Changes in the number of sustained viral responders N in the six scenarios

N	Scenario No.					
	1	2	3	4	5	6
Mean	461	576	692	495	535	572
SD	16	16	17	17	17	17
CV(%)	3.5	2.8	2.5	3.4	3.2	3.0

Similarly, we utilized random numbers obeying the observed frequency distribution of RNA concentration as illustrated in Fig. 2. Patients missing both independent variables did not contribute to the estimation of the number of SVRs, however, in the descriptions that follow, the total count of patients (1,241) in the sample is used in order to provide context for the estimates of SVR, which are based upon patients with at least one predictor available.

The values of the coefficients B_0, B_1, and B_2 used in the computation are summarized in Table 1. The scenario 1 corresponded to the efficacy of past therapy observed among Japanese hemophiliacs (Taki et al. 2004). Scenario 2 and 3 assumed slight and notable improvement in the efficacy, respectively, while the parameters for the viral subtype and concentration were kept unchanged. For scenarios 4 to 6, the improvement in the parameter for either the viral subtype (scenario 4), viral concentration (scenario 5) or both of them (scenario 6) was assumed.

The mean and standard deviations of N after 1,000 times of computation are summarized in Table 2. We observed certain changes in the estimated number of SVRs by changing the scenario, that is, the dependence of the therapeutic efficacy on the predicting variables, as shown in Fig. 3. In the most optimistic scenario,

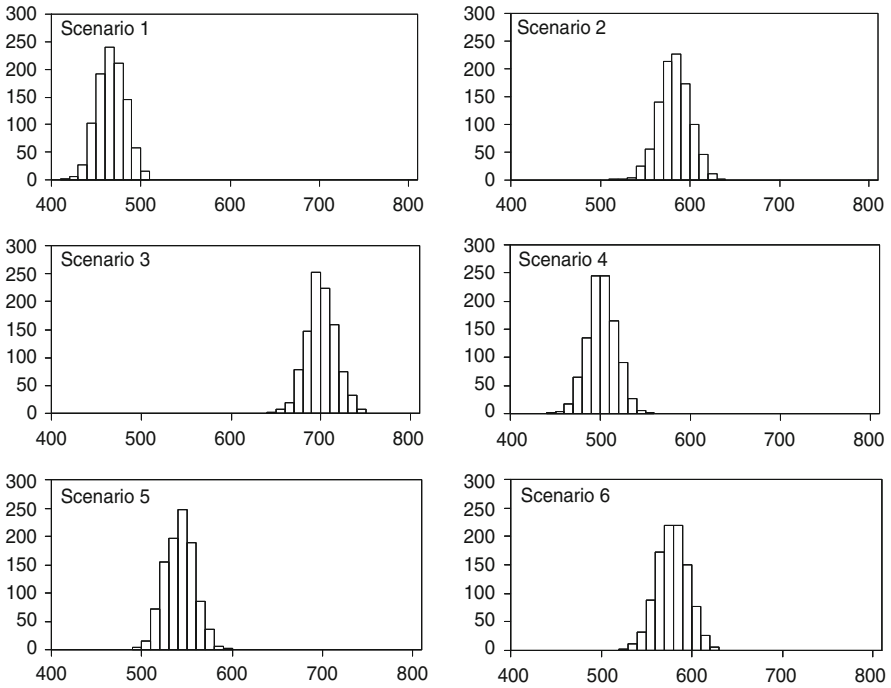


Fig. 3 Changes in frequency distribution of sustained viral responders N in the six scenarios. Horizontal and vertical axes mean values of N and its frequency, respectively

scenario 3, the mean $\pm$ SD of N was 692 $\pm$ 17 (55.8 $\pm$ 1.4%), while in the most pessimistic one (scenario 1), it was 461 $\pm$ 16 (37.1 $\pm$ 1.3%). Therefore, at least 37% of patients will obtain a successful result by receiving the interferon therapy. Values of the coefficient of variation CV arising from the repetition of computations with random numbers were not large and remained in a range smaller than 4%.

5 Discussion

Although the efficacy of the most active therapy using pegylated interferon with ribavirin had been verified (Fried et al. 2002), the various adverse side effects arising from the therapy are preventing patients from starting therapy. The preventive strategy for HCV contamination had been established around 1990, thus more than 18 years has passed since the termination of the possible period of the infection with HCV through coagulation concentrates for hemophiliac treatment. A longer period of the infection, especially one longer than 20 years, is inevitably associated with a higher risk of critical liver disease onset (Lee and Dusheiko 2002), therefore an early decision to start therapy is needed.

According to the present estimates, the faction of SVR is at least 37.1 ± 1.3% in the most pessimistic scenario. This scenario almost corresponds to the efficacy of past therapy using usual interferon with or without ribavirin (Taki et al. 2004). Higher efficacy is expectable in the case of the combination therapy using pegylated interferon and ribavirion (Kontorinis et al. 2005). Thus the estimate from scenario 1 will be the minimum estimate.

In the case of the most optimistic scenario (scenario 3), the fraction of SVR was 55.8 ± 1.4%. The recent observation in 88 Japanese patients with viral serogroup 1 reported that sustained viral response was attained in 47.7% of patients despite the mean of RNA concentration among them was 1,415 kilocopies/mL (Okuse et al. 2008). The value of for a patient with viral subtype 1 is 0.343 (34.3%) when the viral RNA concentration is 1,415 kilocopies/mL in scenario 3. Thus, although under the most optimistic assumption, scenario 3 in fact resembled closely the recent observation. This may encourage patients to receive therapy before the data of their liver function get worse.

Viral subtypes and RNA concentration are prognostic factors for the therapeutic response (Hosogaya et al. 2006), however they are sometimes unknown. This is also preventing a wide and rapid spread of the therapy. The estimated SD originating from the lack of data is not very large and the coefficient of variation was less than 4% in all of the six cases.

Several methods of imputing missing predictor data in logistic regression exercises have been reviewed and examined by Fung and Wrobel (1989), including mean substitution, linear regression of independent variables on each other to substitute linear predictions, dummy variable indicators for missing values, and discriminant function methods. Among those that they examined, the authors favored simple mean substitution and linear discriminant function estimation using available pairs of predictors. Discriminant function analysis assumes multivariate normality among the predictors. One of the predictors in the present investigation is binary, which violates that assumption rendering discriminant function analysis unsuitable. Simple mean substitution is possible in the present investigation, but the single statistic cannot reproduce the distributional profile of the RNA concentration data as well as the numerical method used here does (see Fig. 2). The imputation using random numbers may be easier to understand than the other methods such as the expectation maximizing method (Dempster et al. 1997) among the physicians and patients, because we can demonstrate that the random numbers are actually simulating the observed distribution frequency of the viral subtype and RNA concentration.

When fitting a model, substitution of random numbers that emulate the distribution of an independent variable will ignore any relationship between dependent and independent variables. This will have a tendency to attenuate regression coefficients, which will adversely affect the predictive value of the overall model. This will not, however, be applicable when the relationship between dependent and independent variables is already defined in terms of regression coefficients as in the present investigation (Table 1). Predictions made on the basis of simulations using random variates that distribute in a manner closely following observed values of the independent variables, when used in conjunction with either previously determined

or hypothetical regression coefficients, will have predictive accuracy favorable to that in which, for example, mean substitution is used. Simulations would not be necessarily more accurate than use of only nonmissing independent variable observations, but the simulated predictions would converge to the latter as the number of simulations grows large, which is the reason for choosing 1,000 replications in the present investigation. The reason for using simulation and random number substitution is in order to obtain an index of the precision (Table 2) of the predictions for the patient sample under consideration in the present paper.

Although the present subjects were HIV-negative patients, about 98% of HIV-positive hemophiliacs had been infected with HCV concurrently (Tatsunami et al. 2008). The existence of interaction between HCV and HIV shows that we may have to change the parameters in Table 1 in the case of treatment of HIV-positive patients (Kontorinis et al. 2005).

In any case, usage of the appropriate parameters for the logistic function is necessary for reliable estimates. Thus the determination of parameters among hemophiliacs in the case of the combination therapy using pegylated interferon with ribavirin is needed.

Acknowledgements The present study was supported by KAKENHI (20590521) from Grant-in-AID for Scientific Research (C), The Ministry of Education, Culture, Sports, Science and Technology.

References

Daar, E. S., Lynn, H., Donfield, S., Gomperts, E., O'Brien, S. J., Hilgartner, M. W., Hoots, W. K., Chernoff, D., Arkin, S., Wong, W. Y., Winkler, C. A., & the Hemophilia Growth and Development Study. (2001). Hepatitis C virus load is associated with human immunodeficiency virus type 1 disease progression in hemophiliacs. *Journal of Infectious Disease, 183*, 589–595.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1997). Maximum likelihood from incomplete data via EM algorithm. *Journal of Royal Statitical Society Series B, 39*, 1–38.

Fried, M. W., Shiffman, M. L., Reddy, K. R., Smith, C., Marinos, G., Gonzales, F. L., et al. (2002). Peginterferon alfa-2a plus ribavirin for chronic hepatitis C virus infection. *New England Journal of Medicine, 347*, 975–982.

Fung, K. Y., & Wrobel, B. A. (1989). The treatment of missing values in logistic regression. *Biometrical Journal, 31*, 35–47.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Hosogaya, S., Ozaki, Y., Enomoto, N., & Akahane, Y. (2006). Analysis of prognostic factors in therapeutic responses to thterferon in patients with chronic hepatitis C. *Translational Research, 148*(2), 79–86.

Japanese Foundation for AIDS Prevention. (2008). *Official Report of National Surveillance on Coagulation Disorders in Japan 2007*. Published by Japanese Foundation for AIDS Prevention, Tokyo.

Kontorinis, N., Agarwal, K., & Dieterich, D. T. (2005). Treatment of hepatitis C virus in HIV patients: A review. *AIDS, 19*, S166–S173.

Lee, C., & Dusheiko, G. (2002). The natural history and antiviral treatment of hepatitis C in haemophilia. *Haemophilia, 8*, 322–329.

Okuse, C., Shibuya, A., Itoh, F., Inoue, K., Ohkawa, S., Komatsu, T., Saito, S., Shimizu, H., Suzuki, M., Sekiyama, K., Takatsuka, K., & Tanaka, K. (2008). Efficacy of combination therapy with pegylated interferon and ribavirin in terms of gender, age, and body mass index in patients with chronic hepatitis C belonging to serogroup 1 and with high viral load. *St. Marianna Medical Journal, 36*, 215–227.

Taki, M., Tatsunami, S., Shirahata, A., Fukutake, K., Mimaya, J., & Yamada, K. (2003). Prevalence of hepatitis C virus infection in coagulation disorders in Japan. *International Journal of Hematology, 77*, 528–529.

Taki, M., Tatsunami, S., Mimaya, J., Shirahata, A., Kuwabara, K., Asahara, M., Ohi, T., & Yamada, K. (2004). Factors associating with the efficacy of interferon therapy in Japanese patients with coagulation disorders coinfecetd with HIV and HCV. In the 18th Annual Meeting of the Japanese Society for AIDS. *Journal of AIDS Research, 6*, 40.

Tatsunami, S., Mimaya, J., Shirahata A., et al. (2008). Current status of Japanese HIV-infected patients with coagulation disorders: Coinfection with both HIV and HCV. *International Journal of Hematology, 88*, 304–310.

Virtual High Throughput Screening Using Machine Learning Methods

Cherif Mballo and Vladimir Makarenkov

Abstract Progresses in the field of biotechnology permitted the emergence of an effective screening technique, High Throughput Screening (HTS). In a typical HTS campaign, the main objective consists of the identification of active compounds, called hits. We discuss the possibility of using machine learning methods to predict experimental HTS measurements. Such a virtual HTS analysis will be based on the results of real HTS campaigns carried out with similar compound libraries and similar drug targets. In this way, we analyze an experimental HTS assay from McMaster Data Mining and Docking Competition (Elowe et al. 2005) by means of decision trees, neural networks and support vector machines. First, we study separately the molecular and atomic descriptors in order to establish which of them provide a better discrimination. We present and discuss the results provided by machine learning methods in terms of identification of false positive and false negative hits.

1 Introduction

High Throughput Screening (HTS) is a modern technology intended to automate and accelerate the process of discovery of pharmacologically active chemical compounds (i.e., potential drug candidates). HTS requires testing large numbers of compounds in order to produce a number of hit and lead compounds for future development. Currently, HTS is an integral component of many pharmaceutical, animal health and crop-protection discovery operations. In a typical HTS campaign, the main objective consists of the identification of active compounds (i.e., hits). Several statistical methods have been recently proposed to address the needs of experimental HTS (Brideau et al. 2003; Makarenkov et al. 2007; Malo et al. 2006). Because HTS is a very expensive process, a method allowing one to predict experimental HTS

V. Makarenkov (✉)
Laboratoire de bioinformatique, Departement d'informatique, UQAM, C.P. 8888, Succursale Centre-Ville, Montreal (QC) H3C 3P8 Canada
e-mail: makarenkov.vladimir@uqam.ca

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_56,

measurements *in silico* would be of great benefit for the biotechnological industry. In this study, we analyze a real HTS assay from McMaster University Data Mining and Docking Competition (Elowe et al. 2005) by means of binary decision trees, neural networks and support vector machines. We used the same sampling strategy as that performed in Simmons et al. (2008) in order to compare the methods performances. We also found three other recent studies focusing on the application of machine learning methods in HTS (see Briem and Gunther 2005; Harper and Pickett 2006; Fang et al. 2009). Each of the latter studies was conducted in particular statistical (e.g., sampling strategies) and HTS (e.g., kind of HTS data, proportions of real hits in the data and available descriptors) contexts that were different from ours. The comparisons with the results of these studies could be carried out, but each of them would necessitate the use of a specific sampling strategy, different from that employed by Simmons et al. (2008) and adopted in this article.

2 Data Description

We considered the Test assay (Elowe et al. 2005) from McMaster Data Mining and Docking Competition (http://hts.mcmaster.ca/Downloads). It consists of a screen of compounds meant to inhibit the *Escherichia coli dihydrofolate reductase* (DHFR). Each compound was screened twice: two copies of 625 plates were run through the screening machines. The competition organizers defined as primary hits the compounds (i.e., molecules) that reduced the DHFR activity to 75% of the average residual activity of the high controls. Two lists of hits were published. The first list (consensus hits list) contained all compounds that were classified as hits in both of their replicate measurements. Only 42 out of 50,000 tested compounds were identified as consensus hits. The second list (average hits list) contained 96 compounds, classified as hits when the average value of the two HTS measurements was lower than or equal to 75%. Generally, the proportion of active compounds in real HTS campaigns is around 1%. In the McMaster data set, this proportion corresponds to the experimental measurement values that are lower than or equal to 81.811 (thus, we considered 500 active compounds and 49,500 inactive in our experiments).

Molecular descriptors are numerical values embodying small pieces of chemical information stemming from different molecule representations. Each molecular descriptor takes into account one part of the whole chemical information contained into the real molecule (Todeschini and Consonni 2000). Atomic descriptors are 3D motifs produced from atoms belonging to relevant cavity surfaces (Nebel 2006). Atom pair descriptors were found to be quite effective in modeling (Simmons et al. 2008). Molecular structures were used to compute 3D atom-pair descriptors; such a structure is represented by all of the *atom type – distance – atom type* combinations (Simmons et al. 2008). As examples of atomic and molecular descriptors included in the combined descriptors data set made up in this study we could mention: molecular weight, number of H-accepting and H-donating atoms, number of rotatable bonds, topologic polar surface area and two flavors of log of the octanol/water

partition coefficient (ClogP and SlogP). In this study, we originally considered 209 molecular descriptors and 825 atomic descriptors. The molecular descriptors were computed using the MOE software (Molecular Operating Environment 2008), and the atomic descriptors were obtained using the software supplied by Simmons et al. (2008).

3 Prediction of Experimental HTS Results Using Machine Learning Methods

3.1 Sampling Strategy

In modeling unbalanced data, such as typical HTS data sets, one cannot assess a model performance simply based on the overall accuracy. In this case, the classifiers predict all compounds to have the outcome of the majority class (i.e., inactive compounds in our case) and miss entirely the minority class (i.e., active compounds or hits). As the inactive compounds are dominant (45,500) in our data, we propose to consider all active compounds and sample the inactive ones randomly. Table 1 presents the selected sample sizes and the proportions of the no hit/hit compounds considered in our tests. The notation (*n*:*m*) indicates that, in the sample, we have *n* inactive and *m* active compounds. The ratios reported in Table 1 were selected in order to compare our results with those presented in Simmons et al. (2008).

3.2 Machine Learning Methods

The main goal of HTS is an accurate prediction of active compounds. Thus, HTS is a very natural field to apply predictive data mining techniques such as decision trees, neural networks and support vector machines. In all our experiments, the R2008a version of MATLAB was used to generate the results.

Classification and Regression Trees

Decision trees (Breiman et al. 1984) are very popular in machine learning. A decision tree is a tree-like structure for a set of attributes to be tested in order to predict

Table 1 Ratios of the no hit/hit compounds in the training and test samples

No hit/Hit ratio	Training size	Test size	(%) of Hits
(1:1)	(425:425)	(64:64)	50.0
(2:1)	(850:425)	(128:64)	33.3
(3:1)	(1275:425)	(192:64)	25.0
(4:1)	(1700:425)	(256:64)	20.0
(5:1)	(2125:425)	(320:64)	16.7

the output. In this study, the CART method with the "Classregtree" function and the Gini splitting criterion was used.

Artificial Neural Networks

Artificial Neural Networks (ANN) have been widely used in data mining as a supervised classification technique (Haykin 1999). Larger numbers of neurons in the hidden layer give the network more flexibility. To improve the accuracy, we can also increase the number of epochs (i.e., number of complete passes by the training data set). In this study, we used a backpropagation algorithm. It carries out learning on a multi-layer feed-forward neural network through an iterative process with a set of training samples. For each training sample, the weights were adjusted to minimize the mean squared error between the desired and obtained outputs. We used the MATLAB function "Traingdx" (Adaptative Learning Rate) in this study. The performances were assessed using the mean squared error. The number of epochs was set to 5,000 and the number of neurones in the hidden layer varied from 50 to 250 (with the step of 50), depending of the sample size.

Support Vector Machines (SVM)

Support Vector Machines (SVM) were introduced in Vapnik (1998). They are extensively applied in different fields, including pattern recognition. SVM classification finds a separating hyperplane while maximizing the distance from this hyperplane to the closest data points. The "Svmtrain" training function was used in this study with the linear, polynomial and rbf (radial basis function) kernels to find the separating hyperplane. For the rbf kernel, the scaling factor σ was set to 1. The degree of the polynomial kernel was set to 4.

3.3 Comparison of Molecular and Atomic Descriptors

In HTS, a large number of chemical descriptors is usually available. The selection of a subset of the most predictive variables represents an important issue in this area. The goal of the variable selection is to identify the optimal set of measured variables which best characterize the system under study.

For the considered McMaster data set, we computed the values of 209 molecular descriptors (i.e., variables in our tests) using MOE (Molecular Operating Environment 2008), and the values of 825 atomic descriptors using both MOE and the homemade software supplied by Simmons et al. (2008). For the atomic descriptors, we computed the correlation coefficients between each descriptor and the quantitative response variable (i.e., the biological activity); 209 atomic descriptors associated with the highest values of the correlation coefficient were retained for the further analysis. Thus, both data sets (molecular and atomic) contained the same number of descriptors. We then carried out the CART, ANN and SVM (with linear

and polynomial kernels) methods, separately for the atomic and molecular data sets in order to determine the descriptors providing a better discrimination of the hit/no hit outcome. The proportions of hits and no hits in each considered sample are reported in Table 1 (see also Simmons et al. 2008). In each experiment, we computed the numbers of false positive and false negative hits obtained for each test data set, and then determined the specificity (*Sp*) and sensitivity (*Se*) of the models (Equation 1), where *TP* is the number of true positives, *FN* – number of false negatives, *TN* – number of true negatives and *FP* – number of false positives.

$$Sp = TN/(TN + FP); Se = TP/(TP + FN) \quad (1)$$

Figure 1 illustrates the ROC curves obtained for the atomic and molecular descriptors for each of the considered machine learning methods. When examining the curves in Fig. 1, we can notice that the atomic and molecular descriptors usually yielded very close results in terms of predicting active compounds. However, in the case of SVM (both linear and polynomial), the molecular descriptors were more discriminant than the atomic ones.

Then, we selected the best variables among the molecular and atomic descriptors using the method of "stepwise" variable selection, which is a technique that combines advantages of the forward and backward selections. We used the module "Stepwise" of MATLAB for molecular and atomic descriptors separately; 75 molecular and 64 atomic descriptors were retained. Thus, our new combined data set of predictive variables contained 139 descriptors.

3.4 Results and Discussion

The main objective of this study was to compare the decision tree, neural network and SVM methods in the context of HTS. The values of the 139 descriptors obtained by stepwise selection were used for the final comparison. All presented results are the averages obtained after 100 iterations. At each iteration, we divided randomly each sample into training (85%) and test (15%) subsets to build the model for prediction (see Table 1).

Figure 2 presents the obtained results in terms of percentages of FN, FP, sum of errors (FP + FN) and model sensitivity for the three competing machine learning methods. For the SVM kernel method, we presented the results obtained with the linear and polynomial functions only (the sums of errors were usually higher for the rbf SVM). The SVM method clearly outperformed the CART and NN methods for all measured parameters: FP, FN, FP + FN, and the sensitivity (Fig. 2). For all three methods, one can notice that the recovery of active compounds deteriorates as their ratio in the data set decreases.

Table 2 presents the comparison of our best performances (obtained with the polynomial SVM method) and those obtained for similar sample sizes and training and test sets ratios by Simmons et al. (2008). In the latter paper, 10 different

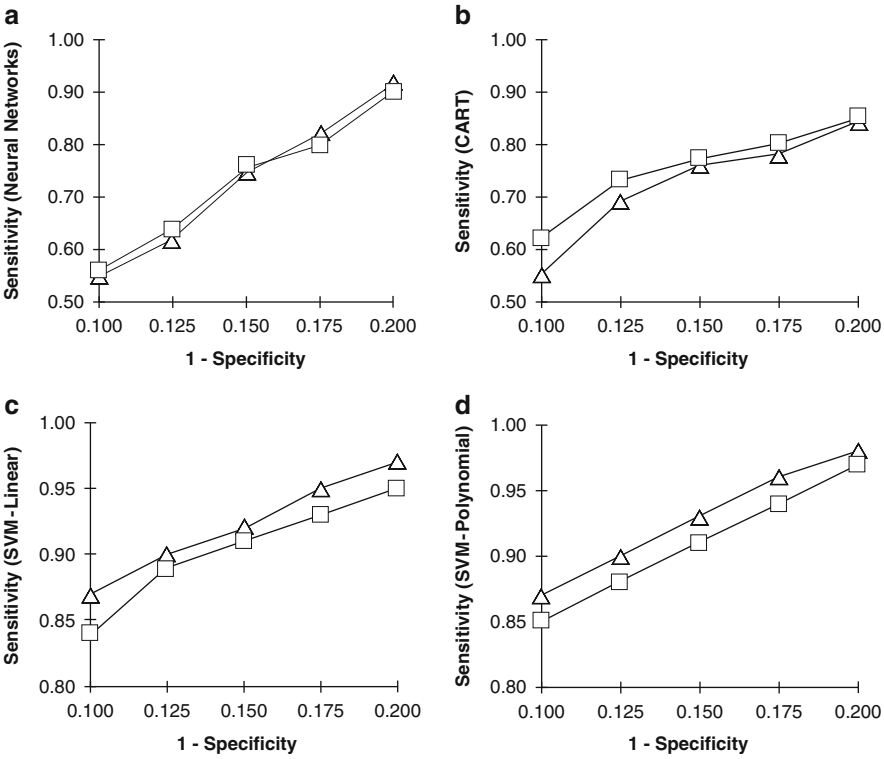


Fig. 1 Molecular versus atomic descriptors comparison (the atomic descriptors are represented by *squares* and the molecular descriptors by *triangles*) for the NN (**a**), CART (**b**), linear SVM (**c**) and polynomial SVM (**d**)

Table 2 The results (polynomial SVM) obtained in this study compared to those provided by Simmons et al. (2008). The latter are indicated between parentheses. The results for ratios (3:1) and (5:1) were not available in Simmons et al. (2008)

No hit/Hit ratio	% FN	% FP	% (FN + FP)
(1:1)	4.30 (16.20)	11.64 (12.00)	15.94 (28.20)
(2:1)	11.17 (23.30)	7.86 (8.20)	19.03 (31.50)
(3:1)	15.59 (na)	6.04 (na)	21.63 (na)
(4:1)	20.75 (15.40)	5.45 (9.60)	26.20 (25.00)
(5:1)	24.77 (na)	5.02 (na)	29.79 (na)

machine learning methods were tested, including InfoEvolve, decision trees (CART and C4.5), oblique decision tree model, kNN (k-Nearest Neighbors), logistic regression, linear discriminant, PLS (Partial Least Squares), NN and FIRM (Simmons et al. 2008). The best overall performances found by Simmons et al. (2008) are reported between parentheses in Table 2. In general, our results were better for all three measured parameters (FN, FP, and FN + FP).

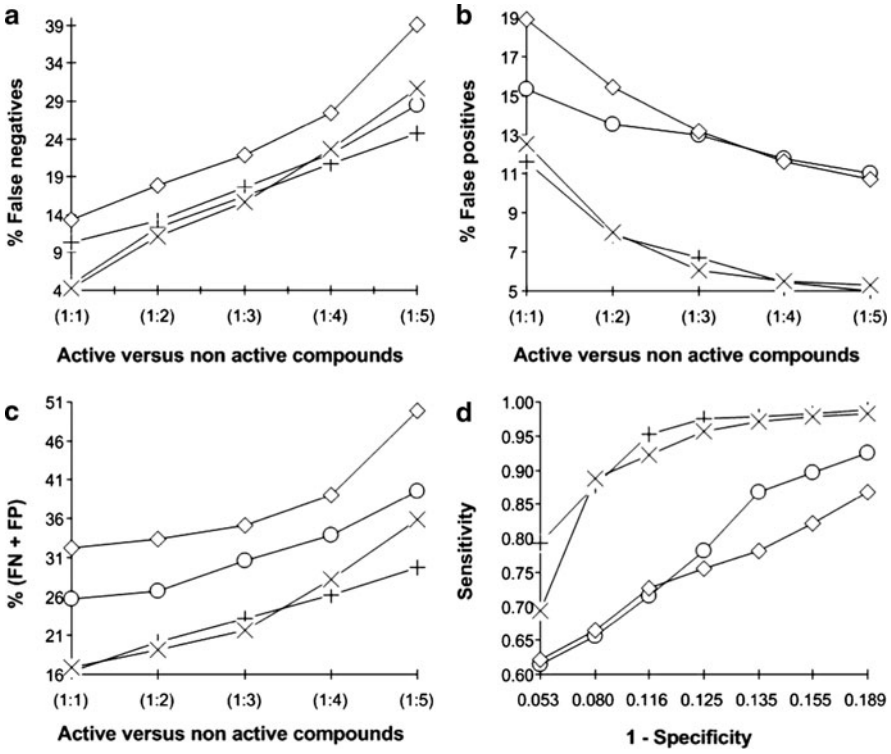


Fig. 2 Comparison of the CART, NN and SVM methods: false negatives (**a**), false positives (**b**), sum of errors (**c**) and ROC curves (**d**). The NN method is represented by *circles*, CART by *squares*, linear SVM by (+) and polynomial SVM by (X)

4 Conclusion and Future Developments

As this study suggests, the machine learning methods can be successfully applied in HTS. As the traditional campaign of identification of hits is a very costly process, an *in silico* method allowing one to predict experimental HTS results accurately would be of great importance. We showed that the molecular descriptors generally had more power to predict active molecules (see Fig. 1c and d). To the best of our knowledge, such a comparison, made in the context of HTS, is novel. Moreover, we carried out a stepwise regression to select the best descriptors in both data sets and create a combined data set of explanatory variables. The results obtained by the three machine learning methods in terms of sensitivity were very encouraging, especially those obtained by the SVM method with linear and polynomial kernel functions. The differences between our results and those obtained by Simmons et al. (2008) should be due to the use of the combined set of atomic and molecular descriptors (Simmons et al. considered only the atomic descriptors) and the application of the SVM method.

As future works, we plan to test other machine learning methods such as kNN and PLS; apply classification methods allowing one to get rid of the descriptors not contributing to clustering (i.e., noisy variables); combine information from computational and combinatorial chemistry (i.e., docking and binding scores) with available descriptors in order to improve the prediction accuracy; and finally, deploy a two-fold machine learning procedure for large HTS data sets having small percentages (1–5%) of hits (such a procedure could use one set of molecular descriptors to select 10–15% of the best samples at the first step, and then perform a second selection within this restricted data set, using a different set of molecular descriptors).

Acknowledgements The authors thank FQRNT for supporting this project, D. Badescu for his help in the analysis of chemical data and K. Simmons for providing us with the software for the computation of the atomic descriptors.

References

Breiman, L., Friedman, J. H., Stone, R. A., & Olshen, C. J. (1984). *Classification and regression trees*. New York: Chapman and Hall.

Brideau, C., Gunter, B., Pikounis, W., & Liaw, A. (2003). Improved statistical methods for hit selection in HTS. *Journal of Biomolecular Screening, 8*, 634–647.

Briem, H., & Gunther, J. (2005). Classifying "Kinase Inhibitor-Likeness" by using machine-learning methods. *ChemBioChem, 6*, 558–566.

Elowe, N. H., Blanchard, J. E., Cechetto, J. D., & Brown, E. D. (2005). Experimental screening of DHFR yields a test set of 50,000 small molecules for a computational data-mining and docking competition. *Journal of Biomolecular Screening, 10*, 653–657.

Fang, J., Dong, Y., Lushington, G. H., Ye, Q. Z., & Georg, G. I. (2009). Support vector machines in HTS data mining: Type I MetAPs inhibition study. *Journal of Biomolecular Screening, 11*(2), 138–144.

Harper, G., & Pickett, S. (2006). Methods for mining HTS data. *Drug Discovery Today, 11*, 694–699.

Haykin, S. (1999). *Neural networks: A comprehensive foundation*. Englewood Cliffs, NJ, U.S.A.: Prentice Hall.

Makarenkov, V., Kevorkov, D., Gagarin, A., Zentilli, P., Malo, N., & Nadon, R. (2007). An effcient method for the detection and elimination of systematic error in HTS. *Bioinformatics*.

Malo, N., Hanley, J. A., Cerquozzi, S., Pelletier, J., & Nadon, R. (2006). Statistical practice in HTS data analysis. *Nature Biotechnology, 24*, 167–175.

Molecular Operating Environment (MOE). (2008). *Chemical Computing Group*, Montreal, Quebec, Canada.

Nebel, J. C. (2006). Generation of 3D templates of active sites of proteins with rigid prosthetic groups. *Bioinformatics, 22*, 1183–1189.

Simmons, K., Kinney, J., Owens, A., Kleier, D., Bloch, K., Argentar, D., Walsh, A., & Vaidyanathan, G. (2008). Comparative study of machine learning and chemometric tools for analysis of in-vivo HTS data. *Journal of Chemical Information and Modeling, 48*, 1663–1668.

Todeschini, R., & Consonni, V. (2000). Handbook of molecular descriptors. London: Wiley-VCH.

Vapnik, V. (1998). *Statistical learning theory*. London: Wiley.

Network Analysis of Works on Clustering and Classification from Web of Science

Nataša Kejžar, Simona Korenjak Černe, and Vladimir Batagelj

Abstract Web of Science (WoS) is a database that provides information about current and past articles published in over 10,000 of the most prestigious, high impact research journals in the world from year 1970 on. A file with full information – records about selected articles – can be downloaded and further analyzed. We collected from WoS complete records on articles from Journal of Classification, articles citing these articles, and articles in WoS cited by them at least 10 times. A special program WoS2Pajek was developed for converting such data into Pajek network files. The citation network between articles, networks of articles × authors, articles × keywords, articles × journals, and the partition according to publication year were obtained from the data. These networks were analyzed in order to identify the most important authors, works and topics that have been involved in the field in the last decades.

1 Introduction

Web of Science (WoS) is an online academic service provided by Thomson Reuters. It provides access to seven world's leading citation databases: Science Citation Index, Social Sciences Citation Index, Arts & Humanities Citation Index, Index Chemicus, Current Chemical Reactions, Conference Proceedings Citation Index: Science, and Conference Proceedings Citation Index: Social Science and Humanities. It covers data on over 10,000 of the highest impact journals of science, technology, social sciences, arts, and humanities, and over 110,000 books and conference proceedings.

WoS allows one to get full information, a record, about an article, a book or other work: its title, authors, abstract, keywords, publication properties (keywords,

N. Kejžar (✉)
Faculty of Social Sciences, University of Ljubljana, Kardeljeva pl. 5, 1000 Ljubljana, Slovenia
e-mail: natasa.kejzar@fdv.uni-lj.si

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_57,

journal, volume, pages, publication year, etc.) and its references. From such bibliographic data many analyses can be done.

Network analysis has been used by Newman (2001) who observed some scientific collaboration networks or, as he called them, acquaintance networks. He analyzed the networks from four different databases of published papers in the 5-year period of 1995–1999 inclusive (MEDLINE, Los Alamos ArXiv, SPIRES and NCSTRL). He discovered some significant statistical differences between prespecified different scientific communities. Collaboration networks were analyzed also for some other data such as Erdős network (Batagelj and Mrvar 2000), boards of directors (On and Balkin 2004), movies database (IMDB) (Ahmed et al. 2007), etc. A wide research of the dynamics of collaboration networks for the fields of mathematics and neuro-science was done by Barabási et al. (2002). They looked at the network from 1991–1998 and investigated the intensity of collaboration, the average separation (in terms of shortest paths) of authors, the clustering coefficients between the fields, as well as through time. They proposed models for the evolution of collaboration networks. The ones predicting connectivity distribution are based on continuum theory, however those that deal with other quantities are studied by Monte Carlo simulations.

Many different network analyses of bibliographic data sets were published in the field of scientometrics. The primary interest of the field is to study science using the scientific methods of science. Citations between scientific works can be studied directly or on agglomerated level as citations between journals or authors. Very informative visualizations of the structure for the whole science (natural and social sciences) from WoS data was constructed by Börner (2009); Börner et al. (2003); Boyack et al. (2005). Visualizations of scientific networks through time and development of various methodological concepts were done by Leydesdorff (see e.g. Leydesdorff et al. 2008; Lucio-Arias and Leydesdorff 2008).

Research was done mainly on 1-mode networks (of collaboration between authors or citations between works or journals). However, as Dorogovtsev and Mendes (2002) pointed out, networks obtained from bibliographic databases are inherently bipartite (2-mode).

In this paper we look at records from WoS database for the field of clustering and classification. We further limit to records from and related to the Journal of Classification (as one of the most important journals in classification) in order to reveal the relevant (groups of) works, authors and topics.

2 Networks from WoS

Initially we intended to analyze the entire field of clustering and classification.

Searches from WoS were done for all years (from 1970–2008) and topics (a) `"cluster analy*"` (67,962 records), (b) `"clustering*"` (49,216), and (c) `"classificat*"` (220,190). Additional search was done for all years and publication name (d) `"Journal of Classification"` and extended with related

works. The results were converted into networks in Pajek (Batagelj and Mrvar 2008) format using program WoS2Pajek (Batagelj 2008).

The usual ISI name of a work (field CR) has the following structure

```
GRANOVET.MS, 1973, AM J SOCIOL, V78, P1360
GRANOVETTER M, 1983, SOCIOLOGICAL THEORY, V1, P203
```

which allows for many inconsistencies. Program WoS2Pajek supports also shorter names (similar to the names used in HistCite (Garfield 2007) output) in the format:

LastNm[:8] + '_' + FirstNm[0] + '(' + PY + ')' + VL + ':' + BP

that eliminate most of the inconsistencies. For example: `GRANOVET_M(1973)78:1360`.

WoS2Pajek produces the following networks:

- citation network **Ci** (stored in file `Cite.net`) of works only
- 2-mode network works × authors **WA** (`WA.net`)
- 2-mode network works × journals **WJ** (`WJ.net`)
- 2-mode network works × keywords **WK** (`WK.net`)
  As keywords are considered regular keywords and also all words from title and abstract without stopwords.

Preliminary network analyses of networks from `"cluster analy*"` showed that the hard core clustering community – members of IFCS (with the exception of some really fundamental works like Ward's *Hierarchical grouping to optimize an objective function* from 1963, or Sneath & Sokal's *Numerical Taxonomy*, 1973) don't play a prominent role in the broad field. Most of the important authors/works, however, belong to the field of biology. This could be due to different publishing cultures in the involved scientific communities (number of coauthors, number of references) or due to the use of the terms cluster, clustering and classification for different meanings.

This was the reason to limit our further analyses in this paper to *JoC data set* which consists of the WoS records on: (a) articles from Journal of Classification (JoC), (b) articles citing these articles, and (c) articles cited at least 10 times from (a or b) articles and having descriptions in WoS.

3 Analyses of Records from JoC

There are 81,581 different works in the JoC data set of which 4,188 have full description records – 599 from JoC. The works come from 9,448 different journals and there are 37,690 authors in the data. Note that for references only the first author is known.

In the original data there was 1 loop (selfreference) in the citation network **Ci**. The inspection of the original paper showed that the error was in the WoS data. We removed the loop from the network and also transformed multiple arcs into single arcs.

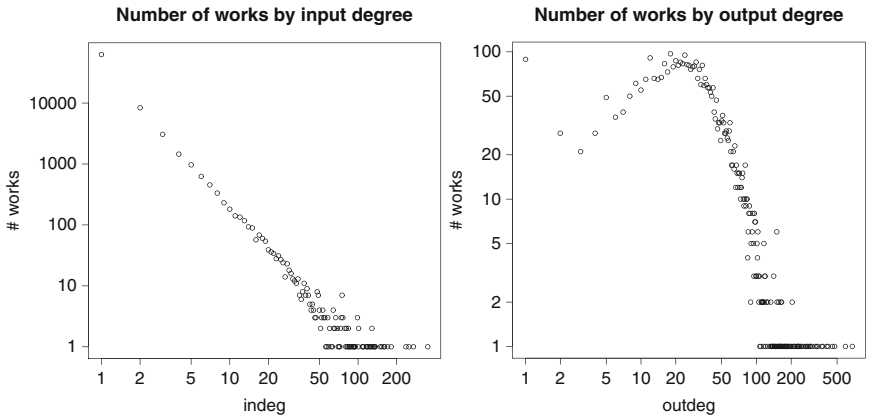


Fig. 1 Number of works by degree

Figure 1 shows the input and output degree distributions for the citation network in log-log scale. The number of works with input degree (number of citations received) decreases very rapidly (power-law). The 15 most cited works with the number of received citations at their beginning are: 349 – HUBERT_L(1985)2:193, 270 – HARTIGAN_J(1975):, 249 – DEMPSTER_A(1977)39:1, 236 – SNEATH_P(1973):, 181 – SCHWARZ_G(1978)6:461, 170 – GOWER_J(1986)3:5, 161 – WARD_J(1963)58:236, 159 – RAND_W(1971)66:846, 153 – JOHNSON_S(1967)32:241, 149 – KAUFMAN_L(1990):, 136 – SAITOU_N(1987)4:406, 134 – JAIN_A(1988):, 134 – MCLACHLA_G(1988):, 132 – KRUSKAL_J(1964)29:1, 129 – ROHLF_F(1999)16:197.

The output degree (number of citations made), however, has a more peculiar shape. It starts high, steps down for 2 and 3, then increases till around 40 and then rapidly decreases. The largest outdegree have works that are either books or overview articles. Note that only works with a full description are considered since only referenced works (without full description) have output degree 0.

Boundary

For further analyses we limit the size of the network (boundary problem) to the works with full descriptions and referenced only works that are referenced often enough – at least k times. We delete vertices for which it holds $(0 < \text{indeg}(v) < k) \wedge (\text{outdeg}(v) = 0)$. In our case we selected $k = 3$.

Frequencies of publications in journals

Let us look at the largest indegrees in the **WJ** network. The journal names in WoS are not unified (normalized) – the same journal can appear under different names. For example: J Roy Stat Soc B, J R Stat Soc B, J Royal Stat Soc B, J Roy Stat Soc B 4, J Roy Stat Soc B Met, J Roy Stat Soc Ser B-Stat Met, J Roy Statist Soc Ser B Metho; or P National Academy S, Proc Nat Acad Sci USA, P

Natl Acad Sci USA; Multivar Behav Res, Multivariate Behav R, Multivariate Behav Res, Multivariate Behavio; J Am Stat Assoc, J Amer Statist Assn, . . .

The list of journals in the bounded network with at least 50 published articles (first number is the number of published articles from the journal) contains: 1009 – J Classif, 425 – Psychometrika, 248 – Syst Biol, 215 – Mol Biol Evol, 207 – Syst Zool, 197 – J Am Stat Assoc, 136 – Comput Stat Data Anal, 120 – Evolution, 117 – Lect Note Comput Sci, 108 – P Natl Acad Sci USA, 104 – Pattern Recogn, 101 – Biometrics, 99 – Bioinformatics, 96 – Multivar Behav Res, 96 – J Mol Evol, 89 – Brit J Math Stat Psy, 88 – IEEE T Pattern Anal, 85 – Cladistics, 82 – Biometrika, 76 – J Roy Stat Soc B, 72 – Science, 71 – J Math Psychol, 70 – Nature, 68 – Math Biosci, 60 – J Marketing Res, 58 – Mol Phylogenet Evol, 56 – Ann Stat, 54 – Genetics, 54 – Discrete Appl Math, 52 – J Theor Biol, 52 – Soc Networks, 51 – Ecology, 51 – Annu Rev Ecol Syst, 51 – Pattern Recogn Lett.

Distribution of articles by the number of authors

The largest number of (co)authors in the articles from JoC is 6 (see Fig. 2), while in other works the number of (co)authors is much larger. Most of the articles from JoC (almost 70%) have only one author, while in other works two authors are more common. This confirms the conjecture that the JoC community has a different publishing 'culture' than the others.

3.1 Collaboration Network

The collaboration network **Co** can be obtained from the 2-mode network **WA** by network multiplication $\mathbf{Co} = \mathbf{WA}^T * \mathbf{WA}$.

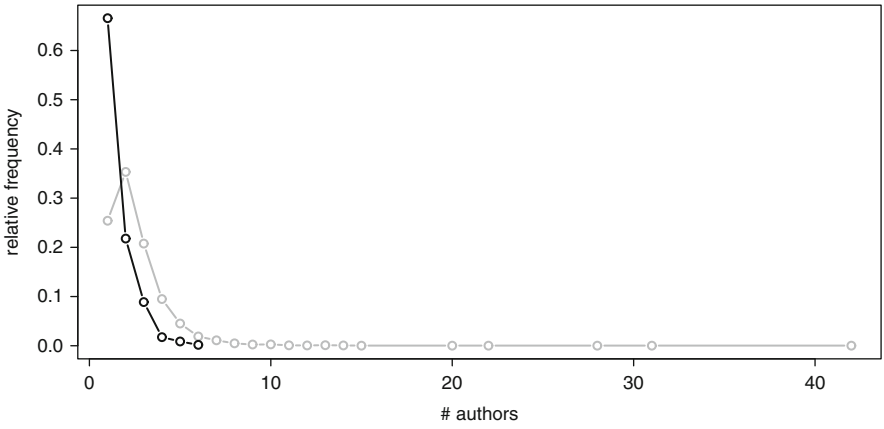


Fig. 2 Distribution of articles by the number of authors (*black* – JoC, *gray* – other)

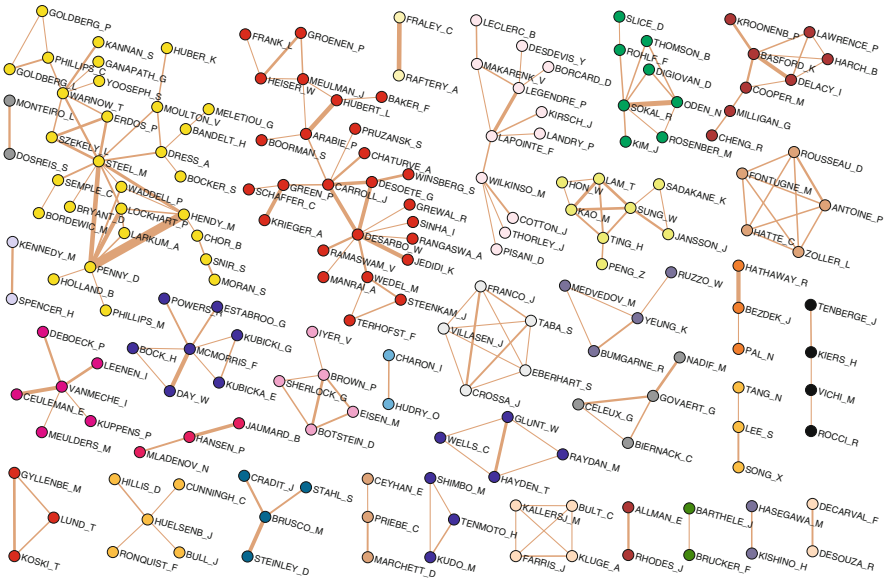


Fig. 3 Main part of line cut at level 3 of collaboration network. Colors represent cuts (connected subnetworks)

In larger collaboration networks we usually try to identify their dense parts using (generalized) cores (Seidman 1983; Batagelj and Zaveršnik 2002). The collaboration network is a sum of cliques determined by each set of authors. For our network the results obtained by standard cores consist mainly of cliques corresponding to the papers with many coauthors. More interesting results are obtained by applying cores to the subnetwork of lines with weight at least 2 or using pS-cores. Even better view on the collaboration structure is obtained by the line cut at level 3 – we preserve in a network only lines with weight at least 3. This network has 298 vertices, 276 lines and 84 components. In Fig. 3 its main part is presented.

3.2 Citation Network Analysis

Main path and CPM path in the citation network

To measure the importance (weight) of arcs in acyclic networks we use the methods proposed by Hummon and Doreian (1989). An efficient algorithm for computing these weights in large networks was developed by Batagelj (2003) and implemented in Pajek by Batagelj and Mrvar. The SPC (Search Path Count) method counts for each arc (u, v) the number of different paths from source (initial vertex) to sink (terminal vertex) passing through it. Therefore the higher the number, more paths pass the arc – more important is the arc. Citation networks are (almost) acyclic. The

problem emerges if there are cycles (nontrivial strong components) in the network. Bounded JoC network has seven such strong components of size 2. All except one are citations between two works from the same publication. We shrink each of them into one vertex.

After weights are computed, the main path and CPM path can be determined. Main path is a path from the source vertex to the sink, starting with the arc with the largest weight and selecting at each step the arc to the neighbors with the largest weight. CPM (Critical Path Method) determines the source-sink path(s) with the largest total sum of weights.

Main path: HOLDER_M(2008)57:814, STEEL_M(2008)57:243, COTTON_J(2007)56: 445, WILKINSO_M(2007)56:330, WILKINSO_M(2005)54:419, EULENSTE_O(2004)53:299, PISANI_D (2002)269:915, PISANI_D(2002)51:151, SEMPLE_C(2000)105:147, SANDERSO_M(1998)13:105, LAPOINTE_F(1997)46:306, PURVIS_A(1995)44:251, BULL_J(1993)42:384, DEQUEIRO_A(1993) 42:368, BARRETT_M(1991)40:486, DONOGHUE_M(1989)20:431, KLUGE_A(1989)38:7, SOKAL_R (1986)17:423, DESOETE_G(1985)2:173, DESOETE_G(1984)1:235, CARROLL_J(1984)1:25, CARROLL_J(1983)48:157, PRUZANSK_S(1982)47:3, #TVERSKY_A(1982)89:123, SHEPARD_R(1980) 210:390, CARROLL_J(1980)31:607, SHEPARD_R(1979)86:87, ARABIE_P(1978)17:21, WHITE_H (1976)81:730, BREIGER_R(1975)12:328, SHEPARD_R(1974)39:373, ARABIE_P(1973)10:148, CARROLL_J(1970)35:283, (HORAN_C(1969)34:139, BLOXOM_B(1968):, CLIFF_N(1968)33:225, MCGEE_V(1968)3:233, YOUNG_F(1967)12:498, ROSS_J(1966)31:27, SHEPARD_R(1966)3:287, TUCKER_L(1966)31:279, WOLD_H(1966):391, TUCKER_L(1964):109, TUCKER_L(1963)28:333, TORGERSO_W(1958):, ECKART_C(1936)1:211).The articles in brackets are all linked to the previous node.

Main topics of the works on the main path are *supertree methods in the consensus setting* in the latest works (mainly published in Systematic Biology), and *multidimensional scaling* in earlier works, published mainly in Journal of Mathematical Psychology and Psychometrika.

CPM path: GOKER_M(2008)8:86, AUCH_A(2006)7:350, THINES_M(2006)110:646, HUSON_D (2006)23:254, DELSUC_F(2005)6:361, GUINDON_S(2003)52:696, CHOR_B(2000)17:1529, STEEL_M(2000)17:839, MAU_B(1999)55:1, RAMBAUT_A(1997)13:235, #MIYAMOTO_M(1995)44:64, HUELSENB_J(1995)44:17, BULL_J(1993)42:384, DEQUEIRO_A(1993)42:368, DOYLE_J(1992) 17:144, PAGE_R(1990)6:119, PAGE_R(1989)5:167, PAGE_R(1988)37:254, PENNY_D(1986)3:403, PENNY_D(1985)34:75, DAY_W(1983)66:97, DAY_W(1983)103:429, ROHLF_F(1982)59:131, ROHLF_F(1981)30:459, #MICKEVIC_M(1981)30:351, SOKAL_R(1981)30:309, SCHUH_R(1980) 29:1, FARRIS_J(1979)28:483, MICKEVIC_M(1978)27:143, MICKEVIC_M(1976)25:260, FARRIS_J (1972)106:645, (FARRIS_J(1970)19:172, FARRIS_J(1970)19:83, KLUGE_A(1969)18:1, ESTABROO_G(1968)21:421, THROCKMO_L(1968)17:355, FARRIS_J(1967)16:44, HENNIG_W (1966):, CAMIN_J(1965)19:311, WILSON_E(1965)14:214, SOKAL_R(1963):).

Main topics on the CPM path are *phylogenetic analysis, evolutionary trees and genome trees* and most of the works on this path are published in Systematic Biology.

Although most works are related with biology, the only common works on both paths are BULL_J(1993)42:384 and DEQUEIRO_A(1993)42:368. All other works are different. Both paths can be found also in the main island in Fig. 4.

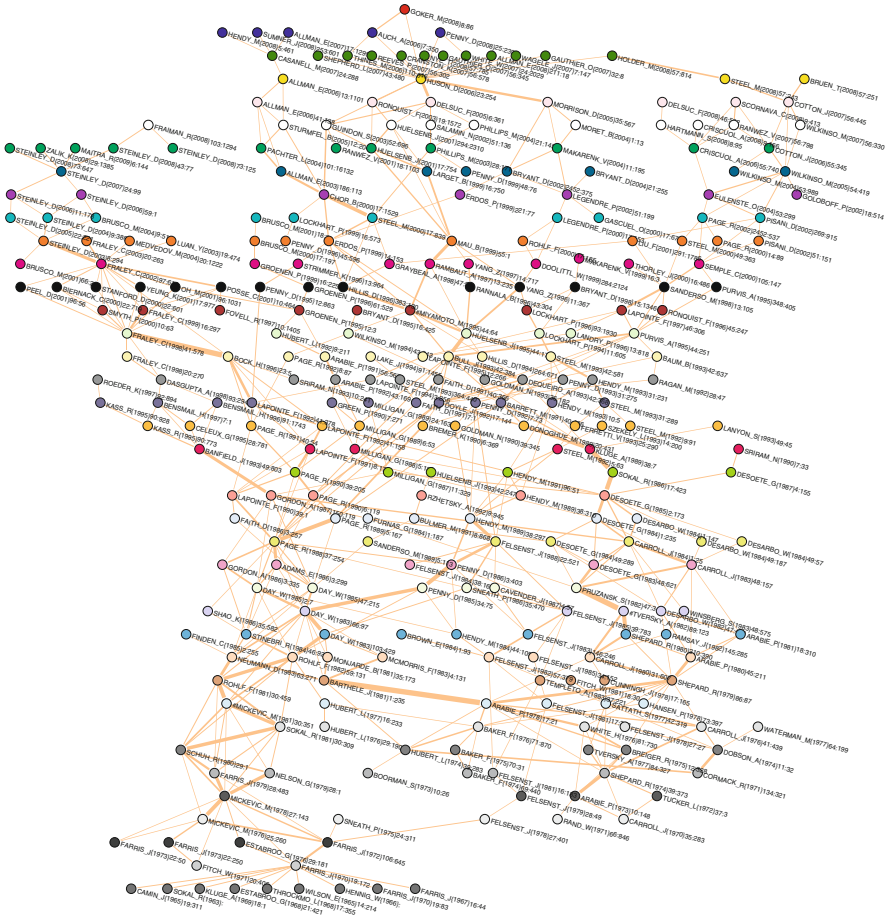


Fig. 4 Main line island of size [5, 300]. Colors represent the same depth level of vertices in the acyclic network

Line islands in citation network

We used line islands to detect subnetworks (clusters) with stronger internal cohesion relatively to its neighbors. A line island of size $[k, K]$ is a weakly connected subnetwork of the selected size in the interval $[k, K]$ where arcs, linking vertices from the island to their neighbors outside island have weights lower than the values of arcs of a spanning tree inside the island (Zaveršnik and Batagelj 2004). Figure 4 represents the largest island of size [5, 300]. All other islands are much smaller.

In this island two main branches can be noticed. The first branch contains the CPM path. It starts with nine works from sixties (Sokal 1963; Farris 1967, 1970, etc.) at the bottom left. Then it follows works about taxonomic studies by Mickevitch and Rohlf, and about consensus index methods by Day, followed by

Penny and Page. From there on, this branch is further divided into two parts: one goes in the middle and further on coincides with the CPM path. The other part (left of it) can be also seen as two branches: one on the far left includes works from Page, Gordon and LaPointe on hierarchical-classification and its applications in the biosciences, continuing with Bock on probabilistic models in cluster analysis, Fraley and Raftery with review of general methodology for model-based clustering, and Steinley on cluster validation. The other branch include works from Milligan with methodology review, Arabie and Brusco on unidimensional and multidimensional scaling, and others, that are related also with works on the second main branch.

The second main branch on the right side of the figure is formed along the main path. It starts with works of Carroll and Arabie. At the bottom, both branches are connected with the strong arc from BARTHELE_J(1981)1:235 to ARABIE_P(1978)17:21. Further, the second main branch continues with works of Shepard, Carroll, DeSarbo, De Soete et al. on multidimensional scaling, additive clustering, tree representation, and meets the first main branch with the article *Partitioning and combining data in phylogenetic analysis* by Bull (BULL_J(1993)42:384), published in Systematic Biology, and with article *For Consensus (sometimes)* by Dequeiroz (DEQUEIRO_A(1993)42:368). After them both branches separate again. The second main branch continues with works of Purvis and Sanderson on phylogenetic supertrees, then splits into two parts: one consisting of the works of Legendre on reticulate evolution, and the other that follows the main path with works from Pisani, Wilkinson and others on combining phylogenetic trees.

3.3 Citations Between Authors

By multiplying $\mathbf{Ca} = \mathbf{WA}^T * \mathbf{Ci} * \mathbf{WA}$, the authors citation network can be obtained. In this authors $\times$ authors network the arc weight corresponds to the number of citations that the first author makes to the second.

3.3.1 Line Islands [10, 400] – Authors Citations

There are 47 simple (one peak) line islands in authors citations network. The largest of them have sizes: 11 – Bezdek, Hathaway, et al.; 10 – Felsenstein, Penny, Hendy, Steel, et al.; 10 – Priebe, Wierman, et al.; 9 – Sokal, Gower, Legendre, et al.; 6 – Maharaj, et al.; 5 – Rohlf, et al. The strongest arcs are in the islands: Brusco $\rightarrow$ Hubert $\leftarrow$ Arabie; DeSarbo $\rightarrow$ Caroll $\leftarrow$ De Soete; and Steinley $\leftarrow$ Milligan. Increasing the upper bound K of island size, the islands with the strongest links join into a single island and the other islands are joining this island. This indicates that there is essentially a single main topic in the network.

Figure 5 presents the largest island where most of the well known names from the IFCS community can be found.

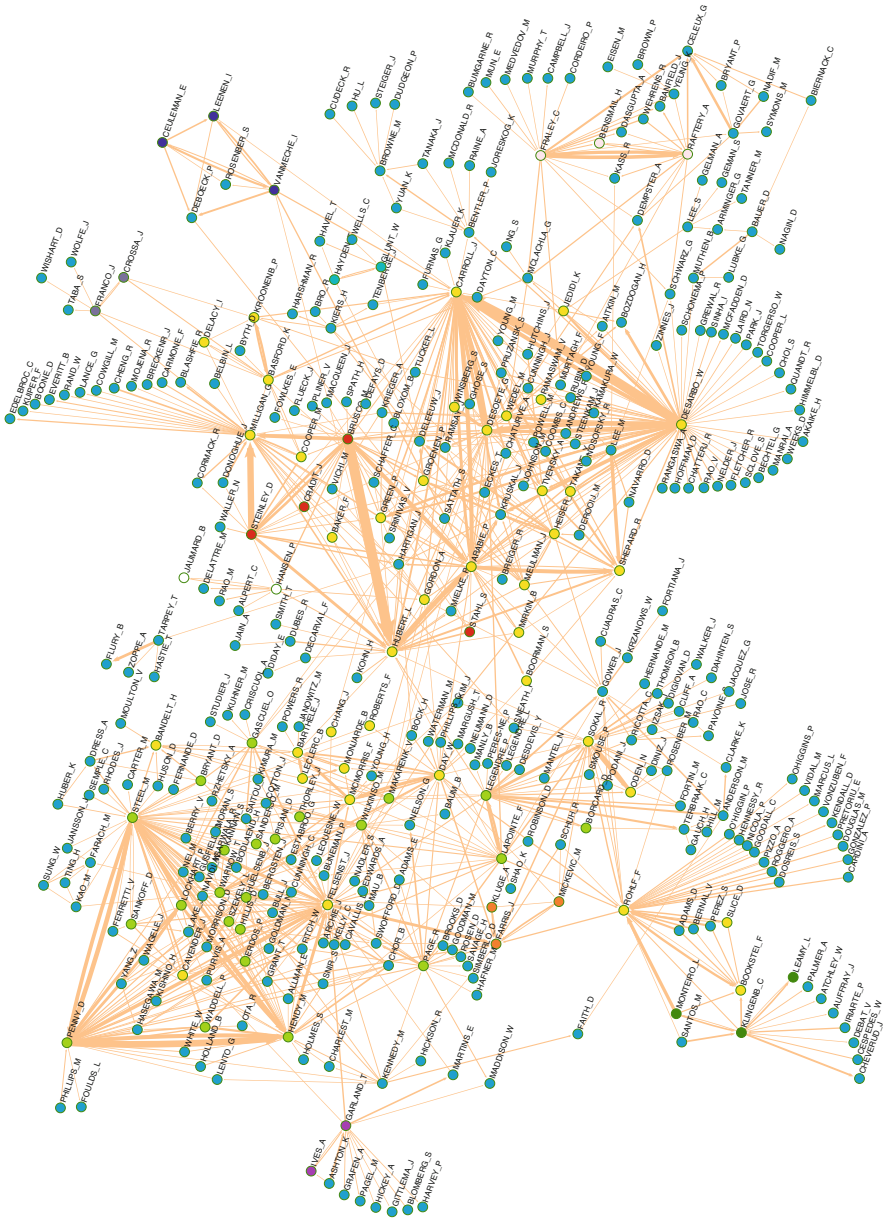


Fig. 5 Largest line island of authors citations

The main groups (clusters) that can be visually identified in the main island can be found also in a part of dendrogram, see Fig. 6, corresponding to hierarchical clustering of vertices of the network using Ward's method on corrected Euclidean distance.

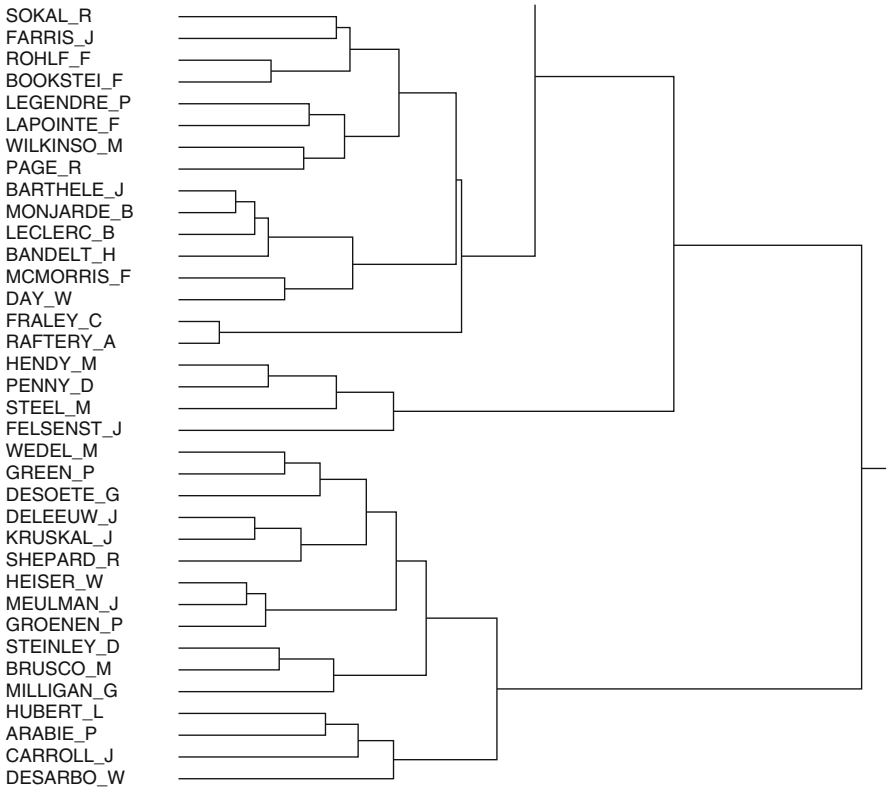


Fig. 6 Part of dendrogram

To find out what is the topic of selected group of authors we considered the 2-mode network $\mathbf{AK} = \mathbf{WA}^T * \mathbf{WK}$. Its arc weight counts how many times the author A used the keyword K. From it we extract the subnetwork group $\times$ keywords and analyze it using methods presented in Ahmed et al. (2007).

4 Conclusion

In the paper we presented the network analysis approach to analysis of bibliographic data. Program WoS2Pajek transforms the original data from Web of Science to a 1-mode citation network and three 2-mode networks (works $\times$ authors, works $\times$ journals, works $\times$ keywords) that can be analyzed separately or in combination with the citation network as derived networks. Using program Pajek we can identify important subnetworks in them and analyze their characteristics.

Because of limited space available for this paper some pictures are rather small and can be read in details only with a magnifying glass. The original color pictures in pdf format can be seen on the web page http://pajek.imfm.si/doku.php?id=examples

References

Ahmed, A., Batagelj, V., Fu, X., Hong, S. H., Merrick, D., & Mrvar, A. (2007). Visualisation and analysis of the Internet movie database. Asia-Pacific Symposium on Visualisation 2007 (IEEE Cat. No. 07EX1615) (pp. 17–24).

Barabási, A. L., Jeong, H., Néda, Z., Ravasi, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A, 331*, 590–614.

Batagelj, V. (2003). *Efficient algorithms for citation network analysis*. http://arxiv.org/abs/cs/0309023. (Submitted on 14 Sep 2003).

Batagelj, V. (2008). *WoS2Pajek*. http://pajek.imfm.si/doku.php?id=wos2pajek. (June 2007).

Batagelj, V., & Mrvar, A. (2000). Some analyses of Erdős collaboration graph. *Social Networks, 22*, 173–186.

Pajek – *Program for large network analysis*. Developed by Batagelj, V., & Mrvar, A., 1996–2010. http://pajek.imfm.si.

Batagelj, V., & Zaveršnik, M. (2002). *Generalized Cores*. http://arxiv.org/abs/cs.DS/0202039. (Submitted on 28 Feb 2002).

Börner, K. (2009). *Atlas of science: Guiding the navigation and management of scholarly knowledge.* Redlands, CA, USA: ESRI Press.

Börner, K., Chen, C., & Boyack, K. (2003). Visualizing knowledge domains. In B. Cronin (Ed.), *Annual review of information science & technology* (Vol. 37, pp. 179–255). Medford, NJ: Information Today, Inc./American Society for Information Science and Technology.

Boyack, K., Klavans, R., & Börner, K. (2005). Visualizing knowledge domains. Mapping the backbone of science. *Scientometrics, 64*(3), 351–374.

Dorogovtsev, S. N., & Mendes, J. F. F. (2002). Evolution of networks. *Advanced Physics, 51*, 566–584.

HistCite: *Bibliometric analysis and visualization software.* Eugene Garfield – President and Founder. HistCite Software LLC, 2007–2010. http://www.histcite.com.

Hummon, N. P., & Doreian, P. (1989). Connectivity in a cition network: The development of DNA theory. *Social Networks, 11*, 39–63.

Leydesdorff, L., Schank, T., Scharnhorst, A., & De Nooy, W. (2008). Animating the development of Social Networks over time using a dynamic extension of multidimensional scaling. *El Profesional de Información, 17*(6).

Lucio-Arias, D., & Leydesdorff, L. (2008). Main-path analysis and path-dependent transitions in HistCite-based historiograms. *Journal of the American Society for Information Science and Technology, 59*(12), 1948–1962.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Scieces USA, 98*, 404–409.

On, J., & Balkin, A. (2004). *They Rule*. http://www.theyrule.net/html/about.php. (Theyrule2004.sql, 21 Nov 2004).

Seidman, S. B. (1983). Network structure and minimum degree. *Social Networks, 5*, 269–287.

Web of Science, Thomson Reuters. http://isiknowledge.com/. (Feb 2009).

Zaveršnik, M., & Batagelj, V. (2004). Islands. Slides from Sunbelt XXIV, Portorož, Slovenia, 12–16 May.

Recommending in Social Tagging Systems Based on Kernelized Multiway Analysis

Alexandros Nanopoulos and Artus Krohn-Grimberghe

Abstract Along with the new opportunities introduced by Web 2.0 and collaborative tagging systems, several challenges have to be addressed too, notably the problem of information overload. Recommender systems are among the most successful approaches for increasing the level of relevant content over the "noise". Traditional recommender systems fail to address the requirements presented in collaborative tagging systems. This paper considers the problem of item recommendation in collaborative tagging systems. It is proposed to model data from collaborative tagging systems with 3-mode tensors, in order to capture the 3-way correlations between users, tags, and items. By applying multi-way analysis, latent semantic correlations are revealed, which help to improve the quality of recommendations. Nevertheless, high-order tensors tend be sparse, a fact that hinders the application of multi-way analysis. To address this problem, we propose the application of kernel-based methods, which act as smoothing functions against sparsity. Experimental comparison, using data from a real collaborative tagging system (Bibsonomy), indicates the superiority of the proposed method against the non kernel-based method and also against other baseline methods.

1 Introduction

Social tagging (a.k.a. collaborative tagging) is a process by which users assign labels in the form of keywords to a set of resources with a purpose to share, discover and recover them. Discovery enables users to find new content of their interest, that is shared by other users. Nowadays, collaborative tagging services proliferate on the Web. Some notable examples are Flickr, Delicious, or My Web 2.0. Collaborative tagging is also very popular for multimedia data, e.g., Last.fm and YouTube.

A. Nanopoulos (✉)
Institute of Computer Science, Information Systems and Machine Learning Lab,
University of Hildesheim, Germany
e-mail: nanopoulos@ismll.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_58,

Along with the new opportunities introduced by Web 2.0 and collaborative tagging systems, several challenges have to be addressed too, notably the problem of information overload. One of the most successful approaches for increasing the level of relevant content over the "noise" lays on recommender systems. Tags can be considered as indicators of interests and preferences. For this reason, tags are a valuable source for recommendation.

Traditional recommender systems fail to address the requirements presented in collaborative tagging systems, because they usually operate over 2-way data arrays, ignoring the 3-way (users, items, tags) aspect of information that is present in collaborative tagging systems. To address the new requirements, recent research, e.g., Hotho et al. (2006), has started to examine novel approaches for developing recommender systems in the context of collaborative tagging systems. Most of these approaches concern the recommendation of tags, in order to create a "collabulary" (collaborated vocabulary). However, the central purpose of Web 2.0 and collaborative tagging systems is to facilitate users in discovering items of interests (e.g., documents, products, songs, video, etc.). Recently, Tso-Sutter et al. (2008) examined the problem of item recommendation in collaborative tagging systems. However, in this work the 3-way-correlation between users, items, and tags is treated as an repeated 2-way-problem.

An effective algorithm for recommending items in collaborative tagging systems models all 3-way correlations between users, items, and tags. This is attained by modelling the data as 3-dimensional matrixes, which are called *3-order tensors*. This approach allows to reveal latent semantics, by performing multi-way analysis based on Tucker decomposition Acar and Yener (2009). Nevertheless, tensors that model social tagging data tend to be highly sparse, because users provide limited numbers of tags. Sparsity can significantly hinder the application of multi-way analysis. To address sparsity, in this paper we propose the use of kernel functions that smooth the data and allow for effective application of multi-way analysis. The superiority of the kernel-based method is supported with experimental results on data from a real-world social tagging system (Bibsonomy).

The rest of this paper is organized as follows. Section 2 summarizes the related work, whereas Sect. 3 reviews the multi-way analysis methods that are employed in the proposed approach. The proposed for providing recommendations is described in Sect. 4, whereas Sect. 5 elaborates on the use of kernels as smoothing functions. Experimental results are given in Sect. 6. Finally, Sect. 7 concludes this paper.

2 Related Work

The problem of recommending tags has attracted significant attention the previous years. Several such algorithms detect conceptual structures in folksonomies similarly to the hyperlink structures detected by search engines. The FolkRank algorithm Hotho et al. (2006) exploits folksonomies by applying the Personalized PageRank algorithm (a modification of global PageRank) to identify important tags to suggest.

Regarding methods that model data from folksonomies with tensors, Xu et al. (2006) proposed a method that recommends tags by using ternary analysis. Symeonidis et al. (2008) used HOSVD for recommending tags. The two aforementioned methods, and all methods summarized in the previous paragraph, focus on how to manage folksonomies for the problem of tag recommendation, whereas the method proposed in this paper is focused on the problem of item recommendation. The two problems are different, because tag recommendation aims at creating a converged vocabulary for tags that will be commonly used and, hopefully, alleviate the main problems in folksomies, like polysemy and synonymy. Conversely, item recommendation is about helping users to find relevant items (e.g., documents, products, songs, video, etc.). Therefore, both problems are interesting and solutions to them can act complementary.

As mentioned, the problem of tag-aware item recommendations has recently started to attract attention. A generic, state-of-the-art item recommendation algorithm is the Tag-aware Fusion, proposed by Tso-Sutter et al. (2008). They propose a generic method that allows tags to be incorporated to traditional, 2-way recommender algorithms, by reducing the 3-way correlations to three 2-way correlations and then applying a fusion method to re-associate these correlations. As will be shown by experimental results, this decomposition breaks the original 3-way structure of the folksonomy and reduces the effectiveness of recommendation.

3 Tensors and Tucker Decomposition

This section provides a concise introduction to the topic of tensors and their decomposition. A *tensor* is a multi-dimensional matrix. A N-order tensor $\mathcal{A}$ is denoted as $\mathcal{A} \in \mathbb{R}^{I_1 \cdots I_N}$, with elements $a_{i_1,\ldots,i_N}$. In this paper, for the purposes of the proposed approach, only 3-order tensors are used. In the following, tensors are denoted by calligraphic uppercase letters (e.g., $\mathcal{A}$, $\mathcal{B}$), matrices by uppercase letters (e.g., A, B), and scalars by lowercase letters (e.g., a, b).

Tucker decomposition for tensor De Lathauwer et al. (2000) generalizes SVD to multi-dimensional matrices. To compute Tucker decomposition on a 3-order tensor $\mathcal{A}$, we need the definition of the following three *matrix unfoldings*:

$$A_1 \in R^{I_1 \times I_2 I_3}, \quad A_2 \in R^{I_2 \times I_1 I_3}, \quad A_3 \in R^{I_1 I_2 \times I_3}$$

Each A_n, $1 \leq n \leq 3$, is called the n-mode matrix unfolding of $\mathcal{A}$ and is computed by arranging the corresponding fibers of $\mathcal{A}$ as columns of A_n. Next, the n-mode product of an N-order tensor $\mathcal{A} \in R^{I_1 \times \cdots \times I_N}$ by a matrix $U \in R^{J_n \times I_n}$ is defined, which is denoted as $\mathcal{A} \times_n U$. The result of the n-mode product is an $(I_1 \times I_2 \times \cdots \times I_{n-1} \times J_n \times I_{n+1} \times \cdots \times I_N)$-tensor, the entries of which are defined as follows:

$$(\mathcal{A} \times_n U)_{i_1 i_2 \ldots i_{n-1} j_n i_{n+1} \ldots i_N} = \sum_{i_n} a_{i_1 i_2 \ldots i_{n-1} i_n i_{n+1} \ldots i_N} u_{j_n i_n} \tag{1}$$

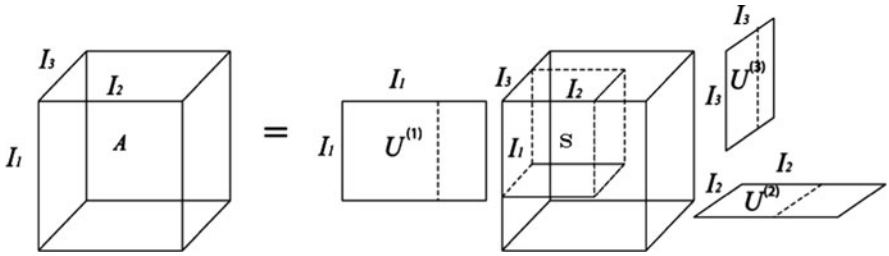


Fig. 1 Visualization of Tucker decomposition

Since the focus is on 3-order tensors, $n \in \{1, 2, 3\}$, only 1-mode, 2-mode, and 3-mode products are being used.

By extending SVD, the Tucker decomposition of 3-order tensor $\mathcal{A}$ can be written as follows De Lathauwer et al. (2000):

$$\mathcal{A} = \mathcal{S} \times_1 U^{(1)} \times_2 U^{(2)} \times_3 U^{(3)} \tag{2}$$

where $U^{(1)}$, $U^{(2)}$, $U^{(3)}$ contain the orthonormal vectors (called the 1-mode, 2-mode and 3-mode singular vectors, respectively) spanning the column space of the A_1, A_2, A_3 matrix unfoldings. S is the core tensor and has the property of all orthogonality. Figure 1 illustrates the Tucker decomposition.

4 Recommendation Based on Tucker Decomposition

This section elaborates on how Tucker decomposition is applied on tensors and on how the recommendation of items is performed according to the detected latent associations. This procedure is summarized with the following sequence of steps.

1. The initial construction of tensor $\mathcal{A}$ From the usage data triplets (user, tag, item), an initial 3-order tensor $\mathcal{A} \in R^{u \times t \times i}$ is constructed, where u, t, i are the numbers of users, tags and items, respectively. Each tensor element measures the preference of a (user u, tag t) pair on an item i.

2. Matrix unfoldings of tensor $\mathcal{A}$ As described, a tensor $\mathcal{A}$ can be matricized i.e., generate matrix representations in which all the column (row) vectors are stacked one after the other. In the proposed approach, the initial tensor $\mathcal{A}$ is matricized in all three modes. Thus, after the unfolding of tensor $\mathcal{A}$ for all three modes, 3 new matrices A_1, A_2, A_3, are created as follows:

$$A_1 \in R^{I_u \times I_t I_i}, \quad A_2 \in R^{I_t \times I_u I_i}, \quad A_3 \in R^{I_u I_t \times I_i}$$

3. Application of SVD on each mode SVD is applied to the three matrix unfoldings A_n, $1 \leq n \leq 3$:

$$A_n = U^{(n)} \cdot \Sigma^{(n)} \cdot V^{(n)^T} \tag{3}$$

4. Computing the low-rank approximations In matrix dimensionality reduction, low-rank approximations are used to filter out the small singular values that introduce "noise". Thus, SVD is truncated to the first c higher singular values and the corresponding singular vectors. The resulting matrix is denoted as rank-c approximation and SVD is optimal in the sense that it computes the rank-c approximation with the minimum *Frobenious norm*.

In the case of tensor dimensionality reduction, a rank-c_1, c_2, c_3 approximation tensor has to be computed, where c_i is the number of dimensions maintained for i-mode. To compute the rank-c_1, c_2, c_3 approximation, c_i singular values and the corresponding left singular vectors from $U^{(i)}$ have to be retained, when applying SVD on the unfolded matrix A_i of i-mode. The selection of c_1, c_2, c_3 determines the final dimensionality of the core tensor $\mathcal{S}$. Since each of the three diagonal singular matrices $S^{(1)}$, $S^{(2)}$, and $S^{(3)}$ are calculated by applying SVD on matrices A_1, A_2 and A_3 respectively, a different c_i value is used for each matrix $U^{(i)}$ $(1 \leq i \leq 3)$. This results to $(U_{c_i}^{(i)})$ matrixes, which denote the c_i-dimensionally reduced $U^{(i)}$ matrix $(1 \leq i \leq 3)$.

5. The core tensor $\mathcal{S}$ construction The core tensor $\mathcal{S}$ governs the interactions among users, items, and tags. Since the dimensions of $U^{(1)}$, $U^{(2)}$ have been selected, and $U^{(3)}$ matrices, the proposed method proceeds to the construction of the core tensor $\mathcal{S}$, as follows:

$$\mathcal{S} = \mathcal{A} \times_1 {U_{c_1}}^{(1)^T} \times_2 {U_{c_2}}^{(2)^T} \times_3 {U_{c_3}}^{(3)^T} \tag{4}$$

where $\mathcal{A}$ is the initial tensor and ${U_{c_n}}^{(n)^T}$ is the transpose of the c_n-dimensionally reduced $U^{(n)}$ matrix, $1 \leq n \leq 3$.

6. The tensor $\hat{\mathcal{A}}$ construction Finally, tensor $\hat{\mathcal{A}}$ is reconstructed by the product of the core tensor $\mathcal{S}$ and the mode products of the three matrices $U^{(1)}$, $U^{(2)}$ and $U^{(3)}$ as follows:

$$\hat{\mathcal{A}} = \mathcal{S} \times_1 {U_{c_1}}^{(1)} \times_2 {U_{c_2}}^{(2)} \times_3 {U_{c_3}}^{(3)}, \tag{5}$$

where $\mathcal{S}$ is the $c_1 \times c_2 \times c_3$ reduced core tensor and ${U_{c_n}}^{(n)}$ is the c_n-dimensionally reduced $U^{(n)}$ matrix, $1 \leq n \leq 3$.

7. The generation of item recommendations The reconstructed tensor $\hat{\mathcal{A}}$ measures the associations among the users, tags and items, so that the elements of $\hat{\mathcal{A}}$ represent quadruplets of the form $\{u, t, i, p\}$, where p is the likeliness that user u will tag item i with tag t. Therefore, items can be recommended to u according to their weights associated with $\{u, t\}$ pair. Specifically, if it is required to recommend to a user u N items that are related to a tag t, then the N items with the highest p values among all $\{u, t, i, p\}$ quadruplets are recommended.

5 Smoothing with Kernel Functions

In Sect. 4 we described that Tucker decomposition applies SVD on the three matrix unfoldings A_n that results to the three matrixes $U^{(n)}$, $1 \leq n \leq 3$, which contain the orthonormal vectors (left singular vectors) for each mode. As already mentioned,

sparsity is a severe problem in 3-dimensional data and it can affect the outcome of SVD. To address this problem, instead of SVD we can apply Kernel-SVD Cristianini and Shawe-Taylor (2004) in the three unfolded matrices. Kernel-SVD is the application of SVD in the Kernel-defined feature space.

For each unfolding A_i $(1 \leq i \leq 3)$ we have to non-linearly map its contents to a higher dimensional space using a mapping function ϕ. Therefore, from each A_i matrix we can derive an F_i matrix, where each element a_{xy} of A_i is mapped to the corresponding element f_{xy} of F_i, i.e., $f_{xy} = \phi(a_{xy})$. Next, we can apply SVD and decompose each F_i as follows:

$$F_i = U^{(i)} S^{(i)} (V^{(i)})^T \tag{6}$$

The resulting $U^{(i)}$ matrixes are then used to construct the core tensor, that is, the procedure continues as described in Sect. 4. Nevertheless, to avoid the explicit computation of F_i, all computations must be done in the form of inner products. In particular, as we are interested to compute only the matrixes with the left-singular vectors, for each mode i we can define a matrix B_i as follows:

$$B_i = F_i F_i^T \tag{7}$$

As B_i is computed using inner products from F_i, we can substitute the computation of inner products with the results of a kernel function. This technique is called the "kernel trick" Cristianini and Shawe-Taylor (2004) and avoids the explicit (and expensive) computation of F_i. As each $U^{(i)}$ and $V^{(i)}$ are orthogonal and each $S^{(i)}$ is diagonal, it easily follows from 6 and 7 that:

$$B_i = (U^{(i)} S^{(i)} (V^{(1)})^T)(U^{(i)} S^{(i)} (V^{(i)})^T)^T = U^{(i)} (S^{(i)})^2 (U^{(i)})^T \tag{8}$$

Therefore, each required $U^{(i)}$ matrix can be computed by diagonalizing each B_i matrix (which is square) and taking its eigen-vectors. Regarding the kernel function, in our experiments we use two widely used functions: (a) the Gaussian kernel $K(x, y) = e^{-\frac{||x-y||^2}{c}}$ (c parameter is called the *width* of the kernel) and (b) the Polynomial kernel $K(x, y) = (x \cdot y + 1)^d$ (d parameter is called the *degree* of the Polynomial).

6 Experimental Results

In this section we experimentally compared four methods: (i) "Kernel", which combines kernel functions with tucker decomposition of tensors, (ii) "Tensor", which uses only Tucker decomposition without kernel functions, (iii) "Epsilon", which simply adds a small ϵ value to every zero element in the tensor (we used $\epsilon = 0.001$), and (iv) "Popular", which recommends to a user the most frequently tagged items that has not been already tagged by him in the training data. Comparison between

Kernel and Tensor helps to understand the usefulness of kernel functions, whereas Epsilon and Popular act as simple baseline methods. The three first methods were implemented using the Tensor Toolbox (csmr.ca.sandia.gov/~tgkolda/TensorToolbox) by setting $c_n = 70\%$, $1 \leq n \leq 3$.

We used real data from the Bibsonomy (www.bibsonomy.org), a social bookmarking system for web resources. The original data was a snapshot taken on April 30, 2007, with 1,037 users, 28,648 items, and 86,563 tags. To remove noise, after applying the 5-core preprocessing (each user, item, and tag must occur in at least 5 posts), 9,763 triples were maintained with 116 users, 361 items, 412 tags. We performed Random Sub-sampling by keeping a fraction of triples for testing (building the tensor) and the rest for training. Each presented results is the average of 30 repetitions (standard deviation is also reported). For each user u - tag t combination in the test data, all methods tried to predict the items tagged by u as t in the test data. We selected recall as the performance measure. Sparsity in the data is characterized by the ratio between the size of the test data to the size of original data, which are given as fractions compared to the size of the original data (e.g., test/training: 0.2/0.8 means that the test and training data was 0.2 and 0.8, respectively, of the original data size, measured in the number of triplets). All measurements are given versus the number of recommended items.

Comparing the results in Fig. 2a and b, we observe that, for the Polynomial kernel ($d = 2$), Kernel is more efficient than Tensor when the ratio between test/training data sizes increases, i.e., when sparsity increases in the results of Fig. 2b compared

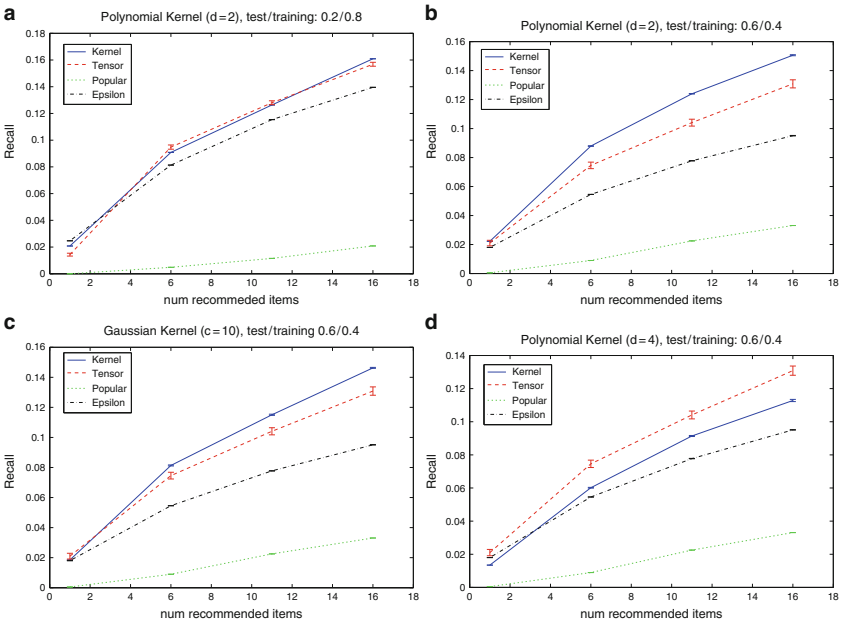


Fig. 2 Recall vs. number of recommended items: (a)–(c) Polynomial kernel, (d) Gaussian kernel

to lower sparsity in the results of Fig. 2a. Please notice that the differences in Fig. 2b are much higher than one standard deviation (depicted with error bars). The same conclusion holds for the Gaussian kernel in Fig. 2c. However, the Polynomial kernel is sensitive to the degree d. Figure 2d present results for $d = 4$. The lower performance of Kernel is due to overfitting that incurs due to high d. We observed that the Gaussian kernel is less sensitive to its width c (result not presented here due to lack of space).

7 Conclusions

We proposed the enhancement of recommending items based on Tucker decomposition with the use of Kernel functions that smooth the tensor data and address the sparsity. Our experimental results with real data presented the superiority of the proposed method over current standard approaches.

Acknowledgements The authors gratefully acknowledge the partial co-funding of their work through the European Commission FP7 project MyMedia (www.mymediaproject.org) under the grant agreement no. 215006. For your inquiries please contact info@mymediaproject.org.

References

Acar, E., & Yener, B. (2009). Unsupervised multiway data analysis: A literature survey. *IEEE Transactions on Knowledge and Data Engineering*, *21*(1):6–20.

Cristianini, N., & Shawe-Taylor, J. (2004). *Kernel methods for pattern analysis.* Cambridge: Cambridge University Press.

De Lathauwer, L., De Moor, B., & Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal of Matrix Analysis and Applications, 21*(4), 1253–1278.

Hotho, A., Jaschke, R., Schmitz, C., & Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. In: *The Semantic Web: Research and applications* (pp. 411–426).

Symeonidis, P., Nanopoulos, A., & Manolopoulos, Y. (2008). Tag recommendations based on tensor dimensionality reduction. In: *2nd ACM Conference in recommender systems (RecSys'2008)* (pp. 43–50).

Tso-Sutter, K. H., Marinho, L. B., & Schmidt-Thieme, L. (2008). Tag-aware recommender systems by fusion of collaborative filtering algorithms. *Proceedings of the 2008 ACM Symposium on Applied Computing*. Fortaleza, Brazil. SAC '08. ACM, New York, NY.

Xu, Y., Zhang, L., & Liu, W. (2006). Cubic analysis of social bookmarking for personalized recommendation. In: *Frontiers of WWW Research and Development – APWeb 2006* (pp 733–738).

Dynamic Population Segmentation in Online Market Monitoring

Norbert Walchhofer, Karl A. Froeschl, Milan Hronsky, and Kurt Hornik

Abstract The objective of the SEMAMO (Semantic Market Monitoring) project is to make use of the increasingly growing information available at Web-based sales and marketing channels for market research, using semi-automatic analysis driven by application domain models. The assumptions are that (i) the Web may serve as a representative "picture" of reality, (ii) the respective online channels map salient market developments, and (iii) all of this accurately and in a timely manner.

Limited server requests and market specific access structures of Web portals inhibit both full scans of sampling populations and random selection of sampled offers. Further, product feature categories entail multiple classifications within offer *clusters* (e.g., geography in tourism). Therefore, SEMAMO proposes an *adaptive* sampling strategy dealing simultaneously with (i) the dynamics of the population frame, (ii) price dynamics, and (iii) multiple (fuzzy) classifications of offered products.

The paper discusses a heuristic method of dynamically segmenting monitored offer populations to stratify online data harvesting depending on both observed price changes and information relevance, and outlines the mechanics of harvest schedule derivation.

1 Introduction

The transparency of e-markets and increasing market dynamics call for more responsive and encompassing approaches towards the monitoring of markets and competition. A natural response to this overall development, advanced information technology (Wen and Wen 2006) – and particularly semantic technologies – provide an unprecedented means to expand both the scope and speed of market observation by reducing the cost of information procurement and, thus, improving competitive decision making. In this respect, SEMAMO (Walchhofer et al. 2009) arguably

N. Walchhofer (✉)
EC3 - e-commerce competence center, Vorlaufstrasse 5/6, 1010 Vienna
e-mail: Norbert.Walchhofer@ec3.at

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_59,

extends the range of current business intelligence methodologies and solutions by designing and implementing a (semi-) automatic market monitoring framework, capitalizing on semantically enriched models of online information extraction.

The ensuing framework is applied to a leading e-commerce domain, tourism, providing (i) a fairly challenging test-bed in terms of market complexity (Werthner and Klein 1999), (ii) an information-rich environment comprising a multitude of online marketing channels, and (iii) structural peculiarities such as volume and access constraints restricting actual data retrieval. Hence, efficient monitoring of online markets depends on a dynamic allocation of access resources, adjusting the observation and analysis effort to varying market conditions.

Section 2 of this paper briefly relates the methodology employed in SEMAMO to preceding work in information retrieval, sampling in online contexts, and modeling of dynamic phenomena such as market prices. Next, Section 3 sketches the supposed model of price dynamics in online markets and introduces the heuristic evidence-based SEMAMO approach towards adaptive market segmentation reflecting the similarity of price change patterns. Referring to the fundamental feedback loop governing the adaptive SEMAMO data harvesting scheme, Section 4 then describes the proposed harvest balancing approach to derive (by combinatorial optimization) feasible data harvest schedules representing both the dynamics and economic utility of market information gathered iteratively. The concluding section presents a summary of the already achieved state of project development, and indicates further project challenges.

2 Related Work

Online market monitoring gathers product information (quality, price) across various online portals by extracting market information (Doorenbos et al. 1997) from heterogeneous semi-structured sources (Wiederhold 1992), using specifically designed wrapper tools (Hammer et al. 1997; Baumgartner et al. 2005). Contrary to many attempts of document retrieval (Appelt 1999) seeking to optimize precision and recall for a wide range of conceivable queries, SEMAMO continually observes a set of deeply structured objects of interest over time.

It turns out empirically that, in many applications, (online) markets typically feature (discrete) jump processes rather than continuously varying prices. Thus, except for its purpose to monitor price dynamics of products within identified markets, the SEMAMO task resembles the tracking of occasional changes in (large) document sets (Cho and Garcia-Molina 2003). Efficiency considerations suggest exploiting evidence of change dynamics for the sake of parsimonious observation; accordingly, SEMAMO capitalizes on an adaptive sampling model reflecting (expected) frequencies of price changes expressed in terms of Poisson-distributed latencies (Grimes and Ford 2008; Matloff 2005). However, the online habitat of SEMAMO inhibits a straight application of proven sampling methods (Levy and Lemenshow 1999); notably, the populations to sample from are explored in a piecemeal fashion as inherent part of the information extraction process proper.

3 Sensor Binning Based on Price Dynamics

By definition, a market monitor aims at tracking price levels as well as their change dynamics of products on offer, for varying degrees of aggregation. Typically, such as in tourism, the offers within a single market are placed on different sales channels – Web portals, in particular – simultaneously; apparently, the same product may appear on multiple portals, with possibly differing prices. Thus, as an analytical unit of observation, it is reasonable to choose an individual product – such as a hotel room, or a package tour, to book – irrespective of multiple portal occurrences, implying that, over time, offer prices may vary reflecting changing market conditions. In what follows, the technical term "sensor" is used to denote a particular offer representing a triplet of (i) channel (Web portal), (ii) market aggregator, and (iii) the individual product on sale as such enabling statistical aggregation over, as well as comparison between, each of these sensor components. However, in an online context, target populations are rather ill-defined because of (i) the pace of change and, as a consequence of this, (ii) the difficulty of actually tracking all population members.

Methodologically, SEMAMO implements a directed data flow from – preselected – Web sources (typically, a set of online stores, or portals, in a given market) of raw online observation data towards aggregated business reports. Starting each harvest cycle with a data harvesting component, consisting of wrappers, and a data transformation unit (Baumgartner et al. 2007) attached, cleansed data is rectified into a regularized representation of price series for the offers monitored. Aligned across data sources in terms of multi-dimensional data warehouse structures, regularized and accumulated price data are ready for a variety of statistical analyses according to customer-defined market reports. Additionally, parameters estimated from accumulated harvest data are used to *adaptively* drive the iterated harvesting of online data.

To illustrate typical price dynamics, the left-hand side of Fig. 1 exhibits a sample of 20 price series taken from the SEMAMO test domain of hotel room offers, tracked over some 40+ harvest cycles. The right-hand part of Fig. 1 contrasts these real price series with simulated ones, using Poisson-distributed jump processes (with $\lambda \sim \gamma(5, 5)$ chosen randomly for each price series, a magnitude random step size $\theta \sim \beta(0.4, 4)$, and the sign of step change also chosen randomly 50:50, in this sequence; starting prices have been generated exponentially distributed with $\lambda = 90$ within the interval $[40, 300]$).

3.1 Harvest Adaptation

Because of technical reasons, access to individual sensors is generally not possible; rather, data wrappers are restricted to formal query binding patterns using a fixed set of filter parameters. This limitation entails a specific kind of cluster sampling. Additionally, very often the feasible number of access operations on a portal at a time is bounded for various reasons such as (i) the amount of requests assigned

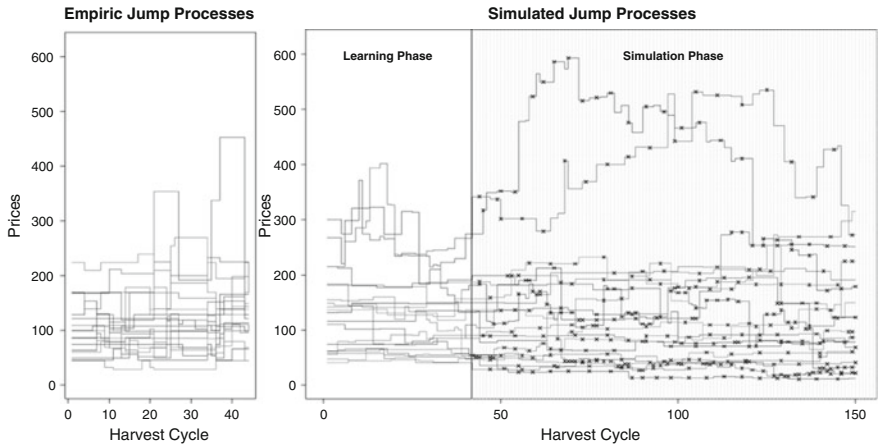


Fig. 1 Empiric and simulated jump processes showing active harvest heuristic's mode of operation by crosses marking update points

might cause an overload of the portal server system and/or may lead to service denial, (ii) the amount of data to harvest might turn out too time consuming, or (iii) the data carry little information suggesting a reduction of observations, or data are not available at all.

Naturally, sensor populations are stratified by the portals monitored. Now, seeking to strike a balance between good population coverage through iterated harvesting and a parsimonious use of portal access resources, an additional stratification of sensor populations based on price change dynamics (Cho and Garcia-Molina 2003) is introduced. To this end, an *active harvest heuristic* estimates the expected change rate, or latency, of a product expressed in terms of regular harvest intervals from both the frequency of observed price changes and the (relative) magnitude of change. Roughly, letting $p_i' = \left|\frac{p_i - p_{i-1}}{p_{i-1}}\right|$, logistically transformed *relative* price changes

$$\varphi(p_t') = \frac{\exp^{\beta(p_t'-\gamma)}}{1+\exp^{\beta(p_t'-\gamma)}} \tag{1}$$

are averaged for estimating the expected latency

$$\tilde{\lambda} = \left\lfloor n \left[\sum_t \varphi(p_t') w(\Delta_t) \right]^{-1} \right\rfloor \tag{2}$$

using the time lapse Δ_t between successive price observations p_{t-1}, p_t of the hitherto observed price series $p_0, p_1, p_2, \ldots, p_t, \ldots, p_n$ of this sensor as a weighting factor (with weighting function w monotonically decreasing in its argument). Then, $\tilde{\lambda}$ is converted to the active harvest weight

$$w_A = \Psi(1-\Psi)^{\tilde{\lambda}} \tag{3}$$

of a sensor, based on some initially set inclusion probability $0 < \Psi \leq 1$. Accordingly, sensors exhibiting more frequent non-negligible price changes receive a higher active harvest weight ($\propto$ probability) for inclusion in the upcoming harvest schedule. Actually, this active weight, adapting to the observation history of a sensor, is combined with a further *delay* weight (compensating sensors overdue for observation because of either recently failed access trials or not having randomly selected them in recent harvest cycles). Active and delay weights are recomputed (updated) every time a harvest iteration has taken place and combined to the *harvest weight* of a sensor.

The right-hand part of Fig. 1 exhibits the heuristically estimated harvest times for the simulated price series by overlaying asterisks to the respective step functions; the heuristic starts working after 40 "observed" harvest cycles used for parameter estimation.

3.2 Dynamic Population Segmentation

Based on the current harvest weights – but regardless of the sensor assignments to portals – a sensor population is segmented into a (pre-defined) number of harvest strata, or "strips", pooling sensors of (adaptively estimated) similar weight, interpreting the stratum centre (e.g., median weight) as sampling rate for randomly selecting sensors of the stratum as observation candidates of the upcoming harvest schedule. Obviously, the effective size of the schedule depends on the respective current stratum sizes of the sensor population, the stratum samples are drawn from. Jointly with the implicit portal stratification, this weight-based segmentation induces a two-way stratification of a sensor population as sketched in Fig. 2 (simplified to

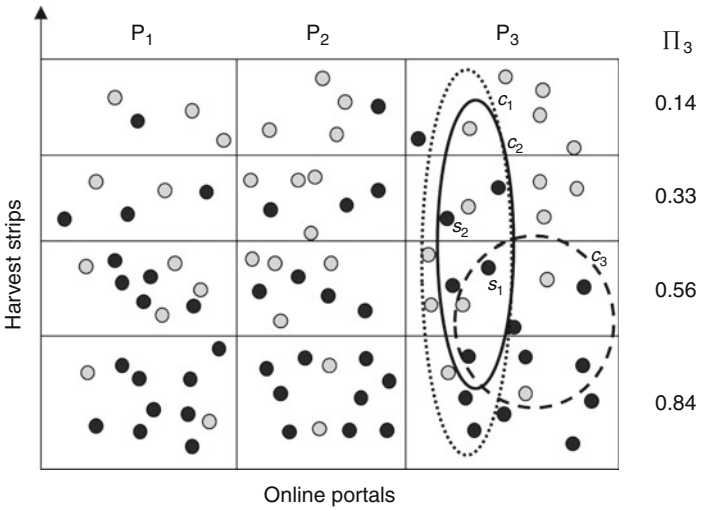


Fig. 2 Sensor stratification over several online portals

three portals and four harvest strips only), with randomly selected sensors in the preliminary pre-selection schedule marked black.

Most of the time, this pre-selection schedule is practically infeasible for actually harvesting online data since portal wrappers can process *certain* query binding patterns only each of which represents, in general, whole sensor classes. By logical necessity, sensor classes are always embedded in portal segments of a sensor population, but usually cut across harvest strips, as indicated to the right of Fig. 2. Worse still, a sensor may entertain multiple class memberships (e.g., because of fuzzy assignments), classes may occur nested (e.g., any 4* hotel is also a 3* hotel by definition), and sensor classes often can be collapsed reasonably into larger sensor classes permitting less complex wrapper queries. Accordingly, the sampled sensors of the pre-selection schedule D have to be mapped into a suitable set of "wrappable" sensor classes to (i) the best degree possible such that (ii) all imposed portal access constraints are met.

4 Harvest Balancing

In what follows, let $s(c)$ denote the function returning the set of sensors in sensor class c and let J_m denote any subset of sensor classes available for building a covering set for D within portal m. Furthermore, assume that $J = \bigcup_m J_m$ still allows the identification of original elements in the J_m sets. Now, assume some real-valued finite cost bound $\chi_m > 0$ for each of the q portals relevant for a given sensor population instance, and let $y_m(c) \geq 0$, $1 \leq m \leq q$, denote the real-valued functions calculating (estimated) access costs of actually scanning the sensor class c on portal m.

4.1 Feasible Harvest Schedules

Using the terminology introduced and writing $|s|$ for the set cardinality of set s, the optimal feasible harvest schedule can be determined as solution of a "0–1 knapsack problem" (Kellerer et al. 2005) as follows:

$$\text{find } \arg\max_J \left| \left(\bigcup_{c \in J} s(c) \right) \cap D \right| = \arg\max_J \sum_m \left| \left(\bigcup_{c \in J_m} s(c) \right) \cap D \right| \tag{4}$$

$$\text{subject to } \sum_{c \in J_m} y_m(c) \leq \chi_m, \, for \; 1 \leq m \leq q. \tag{5}$$

Clearly, the identified solution set J^* of sensor classes may not provide a unique harvest schedule, particularly if $D \subseteq \left(\bigcup_{c \in J} s(c) \right)$.

4.2 Harvest Schedule Tuning

Having obtained a feasible harvest schedule of optimal D-coverage may still leave unsatisfied a couple of additional criteria. Arguably, it is advisable to compose the coverage of D of mutually *disjoint* sensor classes, even though overlaps do not conflict with resource constraints expressed as cost bounds of portals – clearly, the effort of successive processing of harvested data increases proportionally with the size of the data sets generated, regardless of their actual redundancy. Moreover, it is desirable to avoid querying the same offers several times in a single harvest cycle, as this could increase the threat of being recognized as an unsolicited source of Web server load which might cause access denial as a worst case. A further factor entering the objective function could represent a cumulated *valence* based on a non-negative function of the harvest weights of sensors comprised in sensor classes (such as an average, median, or mode value) emphasizing the inclusion of sensor classes contributing a larger share of sensors to sensor population strata with higher sampling rates.

Apparently, in view of the notorious complexity of knapsack problems, solutions exploiting suitably tailored meta-heuristics might work quite efficiently as a replacement of the standard dynamic programming approach.

5 Summary

Transferring market research into online contexts is quite challenging, in particular if one looks for a fairly automatic, application-independent methodology of online market monitoring. This paper has focused on the specific aspect of adaptively generating feasible randomized observation schemes – termed "harvest schedules" – aiming to allocate the data collection effort towards market segments exhibiting higher volatility (in terms of changes in offers as well as, in particular, prices of offers) as compared to apparently more stable market segments. The devised approach towards online data harvesting capitalizes on Web mining methods suggested in the literature, applied to the problem of tracking changes of Web sites or in documents accessible online. In doing so, the specific access conditions of online portals are taken into account such as (i) the re-identification of previously registered online offers (using record linkage techniques), (ii) the update of offer population registries, (iii) computing and maintaining adaptive weights associated which each offer tracked, reflecting the probability of re-harvesting an observation unit in the upcoming harvest schedule, and (iv) the derivation of harvest schedules matched to the constraints of online data access based on combinatorial optimization using adaptively segmented target (sensor) populations. Since routine statistical sampling methodologies do not apply straightforwardly, task-specific heuristics capturing the dynamics of market monitoring have been developed using the domain of (e-)tourism as a prototypical test bed of SEMAMO. The system operates in a cyclic mode, repeating over and over again the job sequence of data

harvesting, data cleansing and rectification, weights updating, and adaptive harvest schedule preparation based on the updated data gathered accumulatively in a historic database.

To date, in the SEMAMO project the processing framework, its main functional and storage components, and the mechanics of the harvest cycle have been developed and implemented prototypically. Parallel to system development, real online data in the domain of tourism are extracted to empirically evaluate the conceived heuristics governing adaptive data harvesting. Currently the proposed method of adaptive harvest scheduling is benchmarked against (i) non-adaptive data extraction schemes, and (ii) adaptive harvesting using a non-stratified random selection of observations, respectively.

References

Appelt, D. E. (1999). Introduction to information extraction. *AI Communications, 12*(3), 161–172.

Baumgartner, R., Frölich, O., Gottlob, G., Harz, P., Herzog, M., & Lehmann, P. (2005). Web data extraction for business intelligence: The Lixto approach. *In Proceedings of BTW*, 48–65.

Baumgartner, R., Frölich, O., & Gottlob, G. (2007). The Lixto systems applications in business intelligence and semantic Web. *ESWC*, 16–26.

Cho, J., & Garcia-Molina, H. (2003). Estimating frequency of change. *ACM Transactions on Internet Technology, 3*(3), 256–290.

Doorenbos, R. B., Etzioni, O., & Weld, D. S. (1997). A scalable comparison-shopping agent for the World-Wide Web. *In Proceedings of the First International Conference on Autonomous Agents*, 39–48.

Grimes, C., & Ford, D. (2008). Estimation of Web page change rates. *JSM'08: Proceedings of the international Joint Statistical Meeting*. Denver, US.

Hammer, J., Garcia-Molina, H., Cho, J., Aranha, R., & Crespo, A. (1997). Extracting semistructured information from the Web. *ACM Proceedings of the Workshop on Management of Semistructured Data*, 18–25. Tucson, US.

Kellerer, H., Pferschy, H. U., & Pisinger, D. (2005). *Knapsack problems.* Berlin, Heidelberg: Springer Verlag.

Levy, P. S., & Lemenshow, S. (1999). *Sampling of populations: Methods and applications.* New York, US: John Wiley.

Matloff, N. (2005). Estimation of internet file-access/modification rates from indirect data. *ACM Transactions on Modeling and Computer Simulation (TOMACS), 15*(3), 233–253.

Walchhofer, N., Froeschl, K. A., Dippelreiter B., Werthner, H., Pöttler, M. (2009). Semamo: An Approach To Semantic Market Monitoring. *Journal of Information Technology & Tourism, 11*(3). 197–210.

Wen, C.-C., & Wen, C.-C. (2006). The use of modern online techniques and mechanisms in market research. *Proceedings of the CIMCA and IAWTIC*. IEEE Computer Society, 93. Los Alamitos, US.

Werthner, H., & Klein, S. (1999). *Information technology and tourism – A challenging relationship.* Wien, New York: Springer Verlag.

Wiederhold, G. (1992). Mediators in the architecture of future information systems. *Computer, 25*(3), 38–49.

Gaining 'Consumer Insights' from Influential Actors in Weblog Networks

Martin Klaus and Ralf Wagner

Abstract Worldwide, weblog users make up a permanently growing conversation database including various private topics, but also discussions about services, products and brands. Hyperlinks create a social network between weblogs in the course of a dialog. This new form of social interaction shifts power in B2C marketing communication toward the consumers.

In this study, we visualize and analyze a social network of weblogs which discuss mobile communication devices. We use different methods of the social network analysis to identify sub-communities and influential weblogs within the whole network. Once these important blogs are identified, we use the netnography procedure to gain "consumer insights" which tell us what the consumers really think and what their needs, wishes, problems and questions are regarding the products.

1 Introduction

Although consumers are concerned about the use and the possible abuse of their transactional information e.g., identity theft or e-mail spamming, they volunteer to disclose non-transactional information by generating and maintaining Web 2.0 contents. This non-transactional information is divided into two qualities: Data mining techniques enable the identification of a person's interests and opinions by exploiting his or her contributions, such as blog entries (Glance et al. 2005). The other quality is provided by the relationship of a person to other persons or groups of persons as well as links to topics. Up to now, marketing research applications failed to grasp both the structure and the contents of non-transactional data. To achieve the best possible courses of marketing action, this paper's new approach combines first a quantitative methodology of social network analysis (SNA) with a qualitative

M. Klaus (✉)
SVI Endowed Chair for International Direct Marketing, DMCC – Dialog Marketing Competence Center, University of Kassel, Germany
e-mail: mklaus@wirtschaft.uni-kassel.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_60,

method of netnography. These two different existing practices from other research areas transferred to marketing matters fit together by first reducing the data and thus the target group to focus on more closely. Then, only the condensed data need to be analyzed qualitatively to gain important information. By seizing the challenge of identifying the relevant knots in a communication network first, and then gaining "consumer insights", this paper aims to:

- propose the ego-network as a useful basic unit of investigation for studies aiming to improve marketing communication.
- demonstrate the usefulness of pre drawing the quantitative SNA to enhance the netnography for gaining "consumer insights".

This paper is structured as follows: Sect. 2 combines the analytic methodology with the qualitative netnography process. First, the SNA, with some of its measures and its relation to the blogosphere, is explained. Netnography is introduced briefly as a method for content analysis. An empirical study of blogs concerning mobile communication is used in Sect. 3 to demonstrate the applicability and relevance of the approach. In the last Section, we summarize our findings and draw conclusions.

2 Methods for Analyzing the Blogosphere

2.1 SNA Measures and Ego Networks

Blogs are interconnected via a huge network (blogosphere). Selected parts of this network concerning topics, brands or products need to be analyzed to make this modern communication channel usable for direct marketing. We aim to identify "important" blogs within the network by social network analysis. The blogosphere is a directed graph where the actors are blogs and the relations are links. Therefore, multiple outgoing edges from one blog to an- other have a meaningful interpretation like permalinks, trackbacks, blogrolls and comments (Klaus and Wagner 2009).

The *degree* centrality $C_d(\cdot)$ provides an impression of the network's structure of the network by considering the number of connections from a given point p_k (Everett and Borgatti 1999):

$$C_d(p_k) = \frac{\sum_{i=1}^{n} a(p_i, p_k)}{(n-1)} \tag{1}$$

where $a(p_i, p_k) = 1$ indicates there is a direkt connection from p_i to p_k and $a(p_i, p_k) = 0$ otherwise n denotes the network size. In the direct marketing context the degree centrality is deemed to be the dimension of possible communication activity within the network. The more links a blog has, the higher is the probability of direct communication with other bloggers. Thus, we assess how applicative blogs are to start canvassing on these blogs with a high degree centrality. The *betweenness* centrality $C_b(\cdot)$ considers the shortest distances within the graph.

$$C_b(p_k) = \frac{2(\sum_{i=1}^{n} \sum_{j=i+1}^{n} b_{ij}(p_k))}{n^2 - 3n + 2} \tag{2}$$

where $i \neq j \neq k$ and $b_{ij}(p_k) = \frac{g_{ij}(p_k)}{g_{ij}}$ with g_{ij} defining the number of geodesics linking p_i and p_j. $g_{ij}(p_k)$ giving the number of geodesics linking p_i and p_j that contain p_k. In this study the betweenness centrality assesses the opportunities for controlling the communication process. If many shortest distances run over a blog, it has a high influence on the network communication, assuming the blogger usually uses the shortest way to communicate. In this way, communication from these blogs can be monitored and assessed by marketers with a view of influencing them as they wish. The *closeness* centrality $C_c(\cdot)$ provides an impression of a blog's centrality in relation to other blogs.

$$C_c(p_k) = \frac{n - 1}{\sum_{i=1}^{n} d(p_i, p_k)} \tag{3}$$

with $d(p_i, p_k)$ denoting the number of edges in the geodesic linking p_i and p_k. In our application domain, the closeness centrality is deemed to be the dimension for independence from other blogs. The measure grows as knots are far apart. Thus this centrality expresses closeness. Consequently, a blog is less dependent on another blog because it has many others close by. Moreover, this measure is assessed as the efficiency of a blog in all the other knots within the network. Considering the distance from one blog to all other blogs in the graph, the closeness centrality indicates how fast a marketing communication measure could spread through the network, starting at blog p_k.

Each blog is also assessed by its *ego network* which comprises a single actor (ego), the actors that are connected to it (alters), and all the links among those alters (Everett and Borgatti 2005). The larger an ego network is, the more alters it has – these alters do not know, or barely know, one another – and the more different the alters are in relation to other criteria, the more powerfully this ego can distribute information. Blogs which seem to be "important" in the examined network because they are surrounded by a strong, dense ego network and show a high centrality measure, are important for marketing purposes because of two facts: (1) They act as multiplicators (Katz and Lazarsfeld 1955) of information within the network, and (2) highly involved bloggers do not often have long actualization breaks on their blogs.

2.2 Netnography

The term netnography was coined by Kozinets (2002). He divides his qualitative methodology into the following four steps which need to be processed. The first two are "Making cultural entre" and "Gathering and analyzing data" which is done in this papers approach by choosing the relevant of the blogosphere, collecting the

data with a web crawler and analyzing it with the above SNA methods. The third step is "Conducting ethical research" which is done by reading the important blogs carefully and extracting consumer insight. Automated text mining methods cannot replace the thorough reading procedure because weblogs have no fixed format, so difficulties in processing these data automatically arise. Particularly, it is almost impossible to indicate and extract the texts in various blog formats precisely. Even if the text data are pre-processed manually it still would be challenging to extract the relevant information with text mining algorithms, because weblogs are written in a colloquial style. Term frequencies and ontologies do not substitute for interpretations The netnography goes deeper in selected details by extracting not only topics or themes from the text but gaining detailed and valued statements and opinions. In line with the argumentation of Kozinets (2006) we conclude that the qualitative way to actually read the weblog content and verbally save the information is best in our application context. In the fourth step, "Providing opportunities for culture member feedback" we develop action alternatives for marketing activities.

3 Empirical Study: Mobile Communication

3.1 Data Description

For this study, the topic *Mobile Communication and Phones* was chosen as an example within the German blogosphere. Therefore, blog URLs related to this topic were collected by searching them via Google.Blog-Search and Technorati with key terms like "Handy", "Mobiltelefon" and "mobile Kommunikation". Not only topic-related blogs in general were searched, but also blogs related to mobile phone brands and products. In all, 173 blog URLs were collected, including 131 generally linked to the topic-related blogs, 18 Apple blogs (iPhone), 15 Nokia blogs, 3 Sony Ericsson blogs, 2 LG blogs, 2 Samsung and 2 XDA blogs. The resulting network is depicted in Fig. 1.

It is clear from the graph that there are many isolates which are not connected to any other blog within the network shown on the edge of the figure. In addition, there are two subnetworks, each with seven blogs, and one with only two blogs. Analysis of these subnetworks indicates that the bigger ones are blog spam farms. Bloggers use these blog spam farms which are interlinked to optimize their search results in engines like Google to get more hits and thus sell their pages for banner marketing. The small subnetwork is just not connected with the rest. After excluding the blog spam farms and isolates from the network, the final blog community network concerning the topic of *Mobile Communication and Phones* includes 112 blogs, and has a density $d(\text{whole net}) = 4.13\%$. This network is the database for all the following analyses.

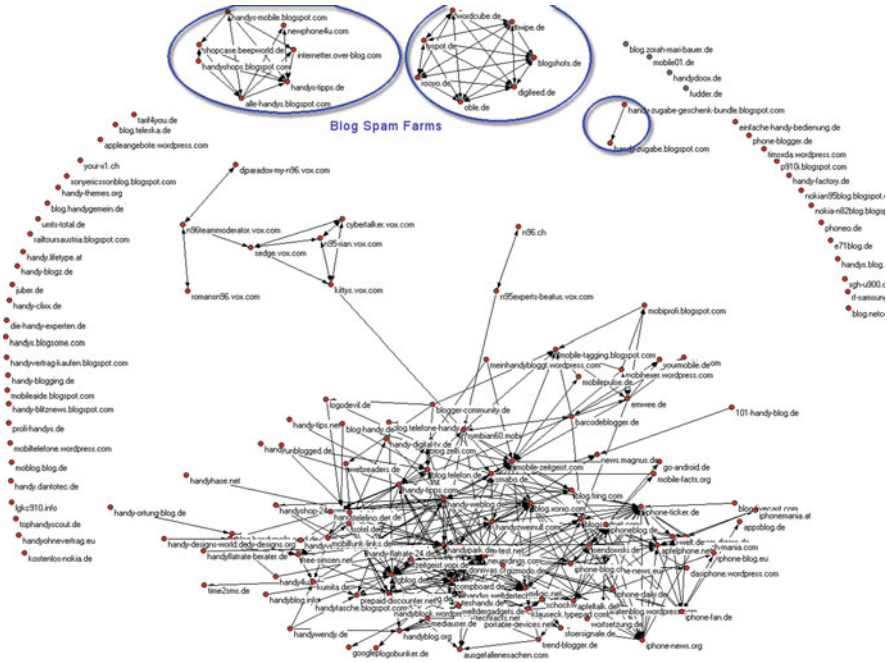


Fig. 1 Blog network with subnetworks and isolates of 173 blogs

Table 1 Extract of the top 112 weblogs with its SNA centrality measures

Blog URL	Type	C_d	C_c	C_b
blog.telefon.de	general Handy	12	42.36	3.90
blog.handymeile-nord.de	general Handy	3	34.79	0.72
guteshandy.de	general Handy	6	38.54	2.13
handy-blogg.de	general Handy	13	40.95	3.36
lgblog.de	**lg**	**27**	**47.63**	**14.01**
iphone-news.eu	apple iphone	5	37.12	1.09
apfelphone.net	apple iphone	10	34.90	0.60
apfeltalk.de	apple iphone	10	36.87	0.88
iphoneblog.de	apple iphone	20	41.41	6.52
⋮	⋮	⋮	⋮	⋮

3.2 SNA Analysis

Obviously, the central and interacting blogs are closer together in the middle of the graph, and less connected blogs fade out toward the border. Also considered for the position of the blogs in the net is the number of the directed links which, for reasons of legibility, are not shown in the graph. The three introduced centrality measures of the 112 blogs are listed for exemplarily chosen URLs in Table 1.

The lgblog.de is highlighted because it has the highest betweenness of all the considered blogs. Thus, this blog appears to be suited to control the communication process. Moreover, lgblog.de has the highest degree centrality within the network, so this blog is also eligible from which to start communication activities. Furthermore, this recommendation is supported by lgblog.de's comparatively high closeness of $C_c(\text{lgblog.de}) = 47.639$. Because of its centrality and independence, this blog provides an efficient knot of origin for marketing communication measures in the blog network under consideration. The ego network of lgblog.de connects with 27 of the other 112 observed blogs, and there are 167 ties in this ego network.

3.3 Netnographical Analysis

Depending on the marketing activity approach adopted, one of the three different centrality measures needs to be used to identify and select the ego for a strong network of combined ego networks.

- Degree: communication activity (participate)
- Betweenness: communication control (monitor)
- Closeness: advertisement (canvass)

For this study, the ego networks of the eight blogs with the high- est betweenness were chosen to be further analyzed with the netnography procedure to monitor the community. The result is a new network with 78 blogs and a density $d(\text{reduced}) = 6.90\%$ which is quite higher than the density of the original network. A higher density stands for a better communication flow (Everett and Borgatti 2005).

Using netnography related to Kozinets (2002), first a "Cultural entre" is needed, which will be the topic-related community from the German blogosphere for this work. Second, the data need to be gathered and analyzed. For this study, the data were crawled and analyzed, and afterwards the "important blogs" concerning mobile communication were identified. Thus, the analysis continues with the third step of "conducting ethical research" by extracting "consumer insights" from the identified network resulting from the eight ego blog networks. All post and comments from the eight ego blogs have been read carefully, and information related to the topic of mobile communication which gave new insights, contained opinions, proposals, ideas or critique were collected in a database.

All detected "consumer insights" were divided into five different insight categories. On the selected blogs, users talked in more detail about hardware, accessories, the combination of applications with hardware, software, and they discussed more general topics. Table 2 gives an overview of all detected "consumer insights" in the five insight categories.

Finally, in the last step of the netnography, opportunities for culture member feedback such as action alternatives and instructions for mobile businesses need to be provided. For this study, the authors provide alternatives and instructions

Table 2 Extract of the top 112 weblogs with its SNA centrality measures

Hardware	Accessories	Apps & Hardware
• Rechargeable battery	• Uniform battery charger	• Medical functions
• "QWERTY" keyboard	• Protection Cover	• GPRS location detection
• ECO Phones	• Engraved designs for phones	• Navigation
• Display resolution	• Bluetooth keyboard	
• RFID technology	• Wireless headphones	
⋮	⋮	
Software	Software	Discussion topics
• Push vs Sync	• Widgets	• Phone prices
• SMS read out	• Screensaver	• Tariff prices
• "Kindle" ebook reader	• SOC software (facebook)	• Customer treatment
• Facebook pics to phone	• TV guide	• Fair news
• Instruction manuals on phones	• Mobile Twitter	• Design & technology dublicates
• iPhone organizer for PC	• P2P call software	
• E-Mail programm	• AppStor Errors	• Product studies

exemplarily related to three different marketing areas. First, the "consumer insights" could be used for product development.

> Example: Consumers often remark that they wish to have standardized ports for mobiles. Thus, recharging, headphones (3.5 klinke adapter), PC connection (mini USB) and memory card-ports should be the same for all mobiles.

In the examined data, the users gave many ideas for new product features and evaluated existing ones. Companies could use this information to have an impression of what customers like, dislike and what their wishes are. By listening closely to customers' needs, companies can improve their image, gain a competitive advantage and finally increase their sales volume.

4 Conclusions and Future Work

Online communities like the blogosphere contain a lot of interesting information for marketing uses. This paper aims to outline methodologies to identify blogs suited for triggering and controlling a marketing communication process on online communities in the blogosphere. The structure of the blog community network related to the topic of mobile communication was crawled and visualized in a first step. In addition, different SNA centrality measures for quantifying the individual blog's position in the communication network have been discussed. The calculation of these measures enables the identification of "important blogs" which have a strong influence on the network. Then, these blogs were examined to gain "consumer insights" about brands, products and topics from which marketing action alternatives were finally derived. The selection of important blogs out of the whole network is an expedient

reduction of burdens in the subsequential netnographical analysis. This emphasises the benefit of supporting the qualitative analysis by both visualisation and quantitative social network analysis. The ego-networks turned to provide suited criteria for the selection of blogs.

Continuing working on this research area, we need to find a decision criterion to decide how many egos should be chosen as "influentials". Moreover, creating two mode networks with additional information describing the actors is likely to provide researchers with even more informative clustering results.

References

Blood, R. (2002). *You've got blog: How weblogs are changing our culture.* Cambridge: Perseus.

Everett, M., & Borgatti, S. P. (1999). The centrality of groups and classes. *Journal of the Mathematical Society, 23*(3), 181–201.

Everett, M., & Borgatti, S. P. (2005). Ego network betweenness. *Social Networks, 27*(1), 31–38.

Glance, N. S., Hurst, M., Nigam, K., Siegler, M., Stockton, R., & Tomokiyo, T. (2005). Deriving marketing intelligence from online discussion. In *Proceedings of KDD* (pp. 419–428). New York: ACM.

Katz, E., & Lazarsfeld, P. F. (1955). *Personal influence.* Glencoe: Free Press.

Klaus, M., & Wagner, R. (2009). Exploring the Interaction Structure of Weblogs, In: Fink, A., Lausen, B., Seidel, W., Ultsch, A. (Eds): *Advances in Data Analysis, Data Handling and Business Intelligence, Berlin*, Springer, 545–552.

Kozinets, R. (2002). The field behind the screen: Using netnography for marketing research in online communities. *Journal of Marketing Research, 39*(1), 61–72.

Kozinets, R. (2006). Definition of netnography. In V. Jupp (Ed.), *Sage dictionary of social research methods.* London: Sage.

Radford, M. (2004). Personal financial services in a digital age. *Journal of Consumer Behaviour, 2*(3), 287–296.

Robertshaw, G. S., & Marr, N. E. (2005). Are abstainers different from voluntary contributors of personal information? Implications for direct marketing practice. *Journal of Direct, Data and Digital Marketing Practice, 7*(1), 18–35.

Thelwall, M. (2004). *Link analysis: An information science approach.* San Diego: Academic Press.

Zheng, Z., & Padmanabhan, B. (2006). Selectively acquiring customer information: A new data acquisition problem and an active learning-based solution. *Management Science, 52*(5), 697–712.

Visualising a Text with a Tree Cloud

Philippe Gambette and Jean Véronis

Abstract Tag clouds have gained popularity over the internet to provide a quick overview of the content of a website or a text. We introduce a new visualisation which displays more information: the tree cloud. Like a word cloud, it shows the most frequent words of the text, where the size reflects the frequency, but the words are arranged on a tree to reflect their semantic proximity according to the text. Such tree clouds help identify the main topics of a document, and even be used for text analysis. We also provide methods to evaluate the quality of the obtained tree cloud, and some key steps of its construction. Our algorithms are implemented in the free software TreeCloud available at http://www.treecloud.org.

1 Introduction

Tag clouds have become very popular on the web. They allow the representation of entire websites in a compact way, through a set of tags whose size or colour reflects their frequency of use (Viégas and Wattenberg 2008). Tags are usually manually associated to the individual articles. However, *word clouds* have been proposed, that can be built directly from a text using the word frequencies, after getting rid of stop words.

The words of a tag or word cloud are often sorted in alphabetical order. This ordering provides no information, although it could be used to express some semantic information on the displayed words, captured using their cooccurrence level. Such improvements of tag clouds have appeared in the literature, for example in Hassan-Montero and Herrero-Solana (2006), where unsupervised clustering, with the number of clusters given as a parameter, is first used to put similar tags on the same line, followed by a reorganization of the lines to group together similar clusters.

P. Gambette (✉)
L.I.R.M.M., UMR CNRS 5506, Université Montpellier 2, France
e-mail: gambette@lirmm.fr

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_61,

Graphs can also be used to display words as well as their cooccurrence relationships, for example in some text mining software like WordMapper, Blake Shaw's visualisation of Del.icio.us tags (Shaw 2005), or Chris Harrison's Bible Visualisation (Harrison 2008). Other approaches have considered multidimensional scaling (van Eck 2005; Fujimura et al. 2008) or factor analysis (Brunet 1993; Viprey 2006) to express the semantic proximity by displaying the most frequent words in two or three dimensions. However, these visualisations are often quite complex, and difficult to read and analyze quickly.

Here we propose to use a tree to reflect the semantic distance between words of a tag cloud, and we call it a *tree cloud*. Then, the distance between two words is given by the length of the path between them in the tree. In fact, the idea of using a tree in this context was already given in Kaser and Lemire (2007), but the tree was not used explicitly and was just a step in an algorithm to display the tag cloud in a compact way.

The problem of finding a tree which reflects a distance matrix was introduced in bioinformatics to reconstruct phylogenetic trees from the information on the distances between their leaves. This very active field has provided algorithms which were also used in text and information processing to represent for example proximity inside a set of texts. It has also been used to reflect the semantic distance between words, according to Google (Cilibrasi and Vitanyi 2007), or inside a text (Brunet 1993; Véronis 2004), but was not used yet to enhance tag clouds.

We describe how to build such tree clouds in Sect. 2. For each step of the algorithm, we give alternative methods or formulas. Then, in Sect. 3, we present some possible test procedures to evaluate the quality of the obtained tree cloud, or the method choosed to generated it. We will focus on a corpus of 138 campaign speeches by Barack Obama, retrieved at http://www.barackobama.com/speeches/.

2 Constructing a Tree Cloud

We consider that we are given an input text containing t words, and detail how to build a tree cloud which describes it.

2.1 Building the List of Frequent Terms

The first step is to extract from this text the list of its most frequent words. Before this process, punctuation should be removed, and other changes in the text can be performed: conversion to lower case or lemmatization (sometimes it should not be applied, see for example "Americans" and "American", which, interestingly, appear in different subtrees in Fig. 1). Some words can also be grouped together, for example different ways to refer to a person: "Barack Obama", "President Obama", "Obama"...

Once the list of most frequent words is obtained, stop words (words unlikely to have a semantic value) may be removed to get a meaningful tree cloud. This

operation is crucial in any word cloud as well, because stop words are among the most frequent, and even on top of the list. Finally, we consider that we obtain a list L of k words, with their frequency.

2.2 Building the Distance Matrix

We then compute some semantic distance between the words in L. Note that we use the word "distance" to refer in fact to a *dissimilarity*, that is, the triangle inequality may not be satisfied; we only guarantee that the distance matrices are symmetric and contain positive numbers, with 0 on the diagonal.

We use the classical principle that the semantic distance between two words in a text is well captured by their cooccurrence. However, there is no ultimate formula to compute the semantic distance, and many have been used in different contexts: more than 20 were gathered and uniformly defined in Evert (2005).

These formulas of cooccurrence distance between two words w_i and w_j are based on a set of portions of the text. $O_{i,j}^{11}$ (resp. $O_{i,j}^{12}$, $O_{i,j}^{21}$, $O_{i,j}^{22}$) counts the number of such portions which contain w_i and w_j (resp. w_i but not w_j, w_j but not w_i, neither w_i nor w_j). A portion can correspond to a sentence, a paragraph, or just a sliding window, depending on the type of text whose tree cloud is being built.

For sliding windows, two parameters have to be chosen: the width w of the window (by default, 30 words), and the size of the sliding step s between two consecutive windows (1 by default). We discuss the choice of these two parameters in Sect. 3.

For the second parameter, a one word sliding step should be chosen to get the most accurate cooccurrence computation. In this case, our algorithm to compute the O^{ab} matrices consists in storing the set L_w of the words of L currently contained in the sliding window with their number of occurrences in the window, updating the content of the O^{11} matrices in $O(min(w, |L|)^2)$, and then updating L_w (in constant time) when the sliding window is shifted. This provides an algorithm of complexity $O(t.min(w, |L|)^2)$ which is in practice much faster than the naive algorithm which computes the cooccurrence for each pair of words in $O(|L|^2.t)$. Note that the beginning of the sliding window of width w starts at position $1-w$, and stops at position t. This ensures that each word has the same weight in the cooccurrence distance.

2.3 Building the Tree

The most popular tree reconstruction algorithm is Neighbor-Joining (Saitou and Nei 1987). For trees reconstructed from textual data, the method mostly used is a variant of AddTree (Sattah and Tversky 1977) proposed by Barthélémy and Luong (1987). It is not clear whether this method is used because it is adequate for such data, or just because Neighbor-Joining was not popular enough when this field of research

started. Other heuristics have been proposed to reconstruct a tree from a distance which is not close to a tree distance, for example a numerical procedure which consists in fitting the distance to a tree distance (Gascuel and Levy 1996) or a more recent one based on quartets (Cilibrasi and Vitanyi 2007).

The bootstrapping methods to evaluate the quality of tree clouds presented in Sect. 3 can help choosing the most appropriate tree reconstruction method for some data. Currently, our program uses only Neighbor-Joining as implemented in SplitsTree, but adding some format conversion functions, to make other tree reconstruction algorithms available, is ongoing work.

2.4 Building the Tree Cloud

The size of keywords can simply reflect the frequency of words, as it is usually the case in tag clouds, or, it can be used differently, for example, to reflect the statistical significance of the various words with respect to a reference corpus. For instance, the words trees representing Barack Obama's and George Bush's discourses could be contrasted that way: the largest words would be those which are the most salient for each of them.

Keyword colours can also convey information. One obvious use is the categorisation according to topics (e.g. Sports, Politics, Business, etc. on a news website). Brightness could be used to show whether the word appears in one same place in the text or in many places (according to some dispersion coefficient). If the corpus is associated with dates, the most recent words can be displayed with the highest intensity or with a different colour, as in Fig. 1.

Information can be conveyed also by edge thickness, length and colour. However, it remains to be seen how much information can be superimposed in the same tree without disturbing its overall readability. A good trick to improve the general aspect of the tree cloud is to force unit edge length. This avoids the long branch problem which occurs with most of the semantic distances: the branches leading to the leaves are very long and the structure of the tree is hidden in the center. The obtained visualisation reflects the semantic distance less faithfully, but the subtree topology appears more clearly.

3 Evaluating the Quality of a Tree Cloud

Tree clouds are useful to get a quick glance at the content of a text. However, one could also use them for further analysis by looking more carefully at clusters of words associated the different subtrees. The tree should then give a good representation of the semantic distance between words in the text, although this distance is just approximated by the tree distance.

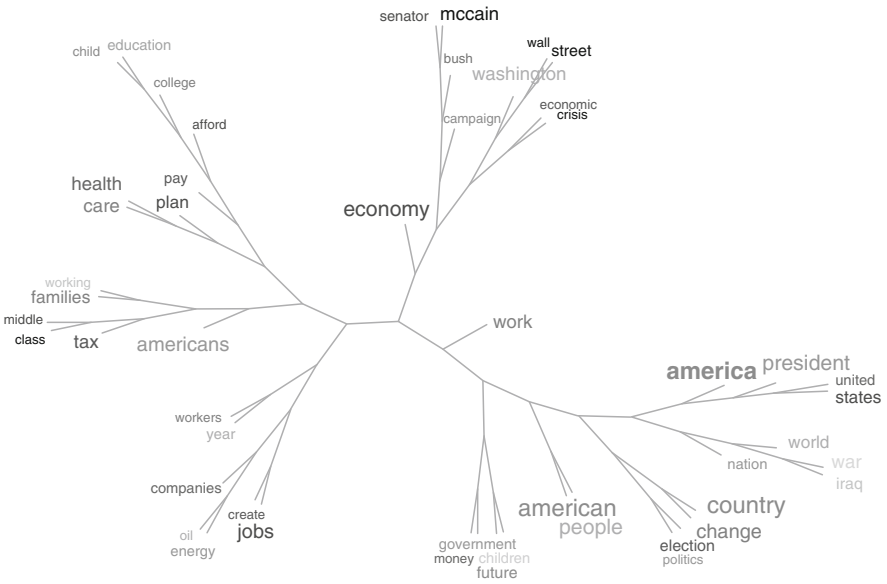


Fig. 1 Fifty word tree cloud of Obama's presidential campaign speeches, with Jaccard distance, and chronology colouring. Light grey corresponds to the beginning of the campaign ("children", "Irak", "war", "world"), while dark grey corresponds to the end ("McCain", "Wall Street", "crisis", "taxes")

In this section, we give some methods to evaluate this quality. They can also be used to choose appropriate distance formulas or tree reconstruction methods for some given input data.

Note that, contrary to phylogenetic tree where the tree distance is supposed to have a biological interpretation (it can represent time, or the number of mutations), the semantic distance formulas which are reflected by the tree have no clear interpretation. In fact, applying an increasing function to the distances would not change their ordering, and the obtained distances can be considered as valid as the input. This explains why the quality of the tree clouds should not be evaluated by direct comparison of the distance matrix and the tree distance.

Instead, we propose an bootstrap evaluation based on the stability of the results. If small changes in the input text provide a similar tree cloud, then it is *stable*, and the method to build it can be considered *robust*. We will also give another criterion, *arboricity*, which evaluates how close the semantic distance is to a tree distance, and discuss how it is related to stability.

3.1 Stability and Robustness

Evaluating the stability of a tree cloud requires two steps: altering the input text, and computing how much the tree has changed. For text alteration, we implemented two procedures: either each word is deleted with probability p, or the text is cut into 100 parts, and some of those parts are removed.

Then, to evaluate stability, we count how many edges of the tree built from the original text are present in the one built from the altered text, seeing each edge as a *split*, i.e. a bipartition of the leaves into two separate clusters. Each edge leading to a leaf is trivially present in both trees, so we neglect those *trivial splits*, and define stability as the proportion of non-trivial splits which appear in both trees.

3.2 Arboricity

Tree reconstruction algorithms are more efficient on distances which fit a tree. Thus, one can expect that the tree cloud will be more stable for semantic distance formulas which provide distances with a good arboricity, that is, close to a tree distance. We first give two formulas which evaluate arboricity, and we will test below whether this can be an objective criterion to evaluate the quality of the tree cloud by avoiding the bootstrap procedure.

The *discrete arboricity*[1] (Guénoche and Garreta 2000) of a symmetric matrix $M \in [0,1]^{n \times n}$ is

$$Arb_d(M) = \frac{1}{\binom{n}{4}} |\{\{i,j,k,l\} \text{ such that } Smax - Smed < Smed - Smin\}|, \quad (1)$$

where *Smin*, *Smed* and *Smax* are the three sums $M_{i,j} + M_{k,l}$, $M_{i,k} + M_{j,l}$ and $M_{i,l} + M_{j,k}$, sorted in increasing order. The *continuous arboricity* (Guénoche and Darlu 2009) of M is

$$Arb_c(M) = \frac{1}{\binom{n}{4}} \sum_{i<j<k<l} \frac{Smed - Smin}{Smax - Smin}. \quad (2)$$

3.3 Distance Comparison on the Obama Corpus

We applied our quality control procedures on the tree clouds obtained on Obama's speeches with the 13 semantic distance formulas implemented in TreeCloud, with text alteration based on removing words with 5% probability.

[1] This formula reflects how much the four point condition, characteristic of tree distances, is verified for each subset of four elements.

Table 1 Average stability (5% alterations) and arboricity of various semantic distances for the tree clouds of the 50 most frequent words of 138 campaign speeches, with sliding windows of size 30 and sliding gap 1

Distance:	Li.	g.m.	Ja.	Di.	m.s.	z.s.	Hy.	χ^2	P.S.	l.l.	od.	NGD	m.i.
Av. stability (%)	56.8	56.8	56.6	56.5	56.0	55.1	55.0	53.9	52.3	51.8	30.9	27.1	17.6
Arb_d (%)	67.2	64.3	65.3	64.4	64.8	66.0	64.4	68.9	53.4	61.8	55.6	59.0	42.1
Arb_c (%)	70.0	66.3	67.6	66.4	66.9	68.2	66.4	72.7	55.4	65.0	55.1	57.5	42.4

Table 2 Average stability of various semantic distances, for the tree clouds of the 50 most frequent words of 138 campaign speeches, to changing the sliding window width (30 by default) or the sliding step (1 by default)

Distance:	Li.	g.m.	Ja.	Di.	m.s.	z.s.	Hy.	χ^2	P.S.	l.l.	od.	NGD	m.i.
$w = 10$	36.2	35.8	35.8	35.5	34.7	35.4	34.6	35.3	34.1	34.4	17.8	10.4	1.2
$w = 100$	28.9	29.1	29.7	29.1	28.5	29.2	25.4	28.4	21.6	27.4	12.5	6.8	1.4
$s = 5$	68.9	70.7	71.0	69.2	69.9	67.0	67.6	61.5	50.1	52.3	27.2	42.1	25.6
$s = 15$	47.9	48.9	48.9	48.5	48.1	48.0	47.8	45.2	39.6	41.6	23.3	24.5	9.7
$s = 30$	34.0	35.0	35.5	35.5	35.6	34.3	33.1	33.9	31.5	32.5	17.7	14.0	3.1

The results are available as supplementary material for this article at http://www.treecloud.org, and a summary given in Table 1 shows that all distance formulas[2] perform approximately equally well, except except mutual information which is very bad, normalized Google distance, and oddsratio. Although oddsratio gives tree clouds with lower stability, it is still an interesting distance, because it provides nice trees even when the edge lengths are not forced to unit length.

The correlation between arboricity and stability is not very good (0.6 correlation coefficient). However, very bad arboricity (below 50%) implies bad stability, and very good arboricity (over 90%) implies good stability.

3.4 Robustness to Parameter Variations

We evaluated stability to decreasing ($w = 10$ words) or increasing ($w = 100$ words) sliding window width, and to variations of the sliding step ($s = 5, 15, 30$) which give similar results for the different distances, shown in Table 2, except for loglikelihood and Poisson-Stirling which seem less robust to sliding step variations.

4 Conclusion

We presented a new visualisation tool which improves word clouds to get a quick overview of the content of a text, as well as some quality control procedures to evaluate how much a tree cloud can be trusted for text analysis. The study of other

[2] The abbreviations correspond to: Liddell, geometric mean (Evert 2005), Jaccard, Dice, minimum sensitivity, z-score, Hyperlex (Véronis 2004), χ^2, Poisson-Stirling, log-likelihood, oddsratio, normalized Google distance (Cilibrasi and Vitanyi 2007), mutual information.

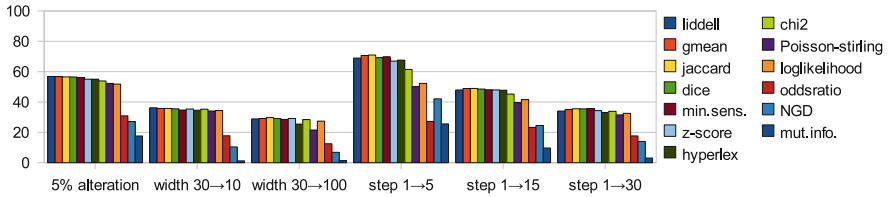


Fig. 2 Visualisation of stability results of Tables 1 and 2

uses in this context (topic-focused tree cloud, tree cloud comparison...) is ongoing work.

Acknowledgements We thank the French ANR project ANR-08-EMER-011-01 (Phyl-ARIANE) for support. We thank Vincent Ranwez and Alain Guénoche for useful discussions, and Virginie Lethier for many references on discourse analysis and lexicometry.

References

Barthélémy, J. P., & Luong, N. X. (1987). Sur la topologie d'un arbre phylogénétique: Aspects théoriques, algorithmes et applications ă l'analyse de données textuelles. *Mathématiques et Sciences Humaines, 100*, 57–80.

Brunet, E. (1993). Un hypertexte statistique: Hyperbase. *JADT 1993*, 1–16.

Cilibrasi, R., & Vitanyi, P. (2007). The google similarity distance. *IEEE/ACM Transactions on Knowledge and Data Engineering, 19*(3), 370–383.

van Eck, N. J. (2005). *Towards Automatic Knwoledge Discovery from Scientific Literature*. MSc Thesis.

Evert, S. (2005). *The Statistics of Word Cooccurrences, Word Pairs and Collocations*. Phd Thesis, pp. 75–91.

Fujimura, K., Fujimura, S., Matsubayashi, T., Yamada, T., & Okuda, H. (2008). Topigraphy: Visualization for Large-scale tag clouds. *WWW2008*, Beijing, China.

Gascuel, O., & Levy, D. (1996). A reduction algorithm for approximating a (nonmetric) dissimilarity by a tree distance. *Journal of Classification, 13*(1), 129–155.

Guénoche, A., & Darlu, P. (2009). TreeOfTrees: A new method to evaluate gene tree distances. Manuscript.

Guénoche, A., & Garreta, H. (2000). Can we have confidence in a tree representation? *Lecture Notes in Computer Science, 2066*, 45–56.

Harrison, C. (2008). *Visualizing the bible*. http://www.chrisharrison.net/projects/bibleviz.

Hassan-Montero, Y., & Herrero-Solana, V. (2006). Improving tag-clouds as visual information retrieval interfaces. *InSciT2006*. Merida, Spain.

Kaser, O. and Lemire, D. (2007). Tag-Cloud Drawing: Algorithms for Cloud Visualization, in Tagging and Metadata for Social Information Organization (workshop at WWW2007), 10 pages, May 2007.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution, 4*, 406–425.

Sattah, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika, 42*, 319–345.

Shaw, B. (2005). Semidefinite embedding applied to visualizing folksonomies. Manuscript, 9 pages, December 2005.

Véronis, J. (2004). Hyperlex, lexical cartography for information retrieval. *Computer, Speech and Language, 18*(3), 223–252.
Viégas, F. B., & Wattenberg, M. (2008). Tag clouds and the case for vernacular visualization. *ACM Interactions, 15*(4), 49–52.
Viprey, J.-M. (2006). Ergonomiser la visualisation AFC dans un environnement d'Exploration textuelle : une projection "Géodésique". *JADT 2006*, 981–992.

A Tree Kernel Based on Classification and Citation Data to Analyse Patent Documents

Markus Arndt and Ulrich Arndt

Abstract We consider the problem of representing patent documents in such a way that a kernel matrix reflecting the similarities of the documents can be efficiently computed.

The European classification system ECLA is a deep level hierarchical taxonomy comprising about 130,000 classification symbols. Depending on their technical content, patent documents are assigned one or more ECLA classification symbols. In this study we represent the complete ECLA taxonomy as a tree labelled by the classification symbols, called the ECLA tree. Within the ECLA tree a positive value is attached to each node of the tree reflecting the technical specificity of the corresponding classification symbol. Based on the directly assigned symbols as well as on symbols of the cited and citing documents, patent documents are mapped to subtrees of the ECLA tree. Taking into account the specificity of the tree nodes, we define an inner product on subtrees representing the documents. It is shown that the inner product is a valid kernel function which can be effectively used for discovering clusters in a set of patent documents.

1 Introduction

Defining similarity measures based on taxonomies has been an active area of research for some time. Several such measures have been developed in the context of WordNet in order to calculate the semantic similarity between expressions (Lin 1997; Resnik 1999).

In this study we focus on the European Classification system **ECLA** which is a taxonomic scheme for classifying technical documents, in particular patent documents (European classification 2008; Dickens 1994). When comparing two documents each with one or more ECLA classification symbols assigned we want to answer the question of how technically related these documents are. Moreover,

M. Arndt (✉)
European Patent Office, Erhardt Street 27, 80649 Munich, Germany
e-mail: marndt@epo.org

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_62,

as patent citations clearly indicate that the citing and cited document share technical content, a similarity measure for patent documents should include the citation information (Li et al. 2007).

A kernel function, which expresses the relatedness of documents, is developed by combining classification and citation information such that use can be made of the **kernel toolbox** for clustering patent documents, for calculating distances, for visualization document sets, etc. (Shawe-Taylor and Cristianini 2004). The defined kernel function has been applied to cluster a set of patent documents in order to validate its usability by way of an example.

2 European Classification System ECLA

As an extension to the International Patent Classification system (International Patent Classification IPC) the European Classification system ECLA is a deep level hierarchical taxonomy comprising about 130,000 classification symbols (Dickens 1994; European classification 2008). It is organized in nine sections designated by one of the capital letters A through H and Y. By introducing a generic root symbol (ECLA Root) having edges to the nine sections, one can represent the complete ECLA taxonomy as a tree labelled by the classification symbols. This tree is called the ECLA tree.

In a manually performed classification process, a document is assigned to one or many ECLA symbols when its technical content matches the definition of one of the ECLA symbols. As a general rule, a document is given the classification symbol at the deepest appropriate hierarchical level which is not necessarily a leaf of the classification tree. Formally, document classification can be regarded as a mapping from the set of documents to the power set of classification symbols.

The technical specificity of an ECLA symbol Θ can be quantified as follows: Let S_Θ be the subtree rooted at the symbol Θ and let $|S_\Theta|$ denote the size of S_Θ which equals its number of nodes (cf. Fig. 1), then

$$\text{Specificity}(\Theta) = \frac{1}{1 + \ln |S_\Theta|} \tag{1}$$

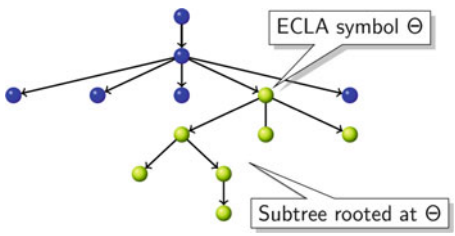


Fig. 1 ECLA specificity

By this definition, a leaf symbol has the specificity value of 1; and the specificity of an ECLA symbol within the tree becomes smaller the larger the subtree attached to that symbol. The rationale behind this definition is that the specificity of an ECLA symbol can be regarded as the quantity of domain specific information contained in the symbol. As a leaf node characterizes a technical concept most precisely, it represents the *relative maximum* information with respect to the associated technical domain. The specificity of a node within the ECLA tree depends on the structure of the subtree which is rooted at this node. When an ECLA symbol has a large subtree, this indicates that there is much domain specific information attached. Hence, this node contains *relatively general* information and, consequently, the specificity associated with that node must be smaller.

3 Patent Citations

Search reports are established for a patent application in order to identify prior art documents relevant to the claimed subject-matter. During the examination phase which follows the search, the assessment of novelty and inventive step of the subject-matter defined in the claims is based on the documents which have been cited in the search report. For these reasons, patent citations provide a strong link from the technical content of the application to the technical content of other documents. In this study we make use of this link by relating cited and citing documents via their ECLA symbols, a concept which has previously been applied by Li et al. (2007).

4 Tree Kernel

A standard model for representing text documents is the vector space model (VSM) which uses vectors to represent documents. According to the VSM, each dimension of the vectors corresponds to a particular word. If a word occurs in a document, its value in the corresponding vector is non-zero such that the bag of words of a document is mapped to a vector (Salton et al. 1975).

In our model we make use of the directly assigned ECLA symbols as well as of symbols of the cited and citing documents to describe the information content of patent documents, whereby the hierarchical ECLA taxonomy suggests that documents be represented by rooted subtrees of the ECLA tree (cf. Fig. 2). We further propose that these subtrees be mapped to vectors δ, whereby each dimension of the vectors corresponds to a particular ECLA symbol. A component of a vector is assigned a non-zero value whenever the associated ECLA symbol is part of the ECLA subtree which represents a particular document. The function which determines the value is related to the specificity of the respective ECLA symbol (cf. Sect. 2 and (2)). The kernel function which captures the technical content two documents have in common is than given by the inner product between the feature vectors (Shawe-Taylor and Cristianini 2004): $\kappa(\delta_A, \delta_B) = \langle \delta_A, \delta_B \rangle = \delta_A^T \delta_B$.

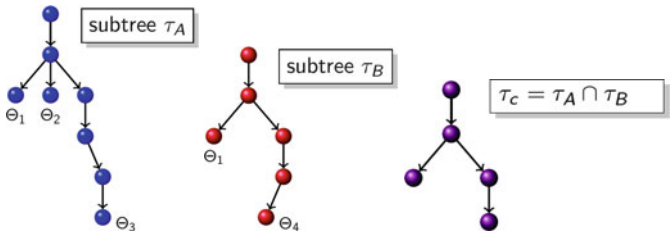


Fig. 2 The subtree τ_A represents a document A with three ECLA symbols Θ_1, Θ_2 and Θ_3; the other subtree τ_B a document B with two ECLA symbols Θ_1 and Θ_4. The common technical content of the two documents A and B is expressed by the largest common subtree τ_c of the trees τ_A and τ_B. Thus, τ_c equals $\tau_A \cap \tau_B$

As in the standard bag-of-word vector space model we represent a set of documents by a document matrix E. In this matrix each row is related to a document and each column to an ECLA symbol. Naturally, the document matrix E is a sparse matrix, since the ECLA subtree which represents a document comprises just a small number of all the ECLA symbols. The kernel or Gram matrix which expresses the relateness between any two documents in the set is then the document by document matrix $K = EE^T$.

Four data matrices are defined in order to effectively calculate the document matrix E: a document classification matrix D_{ec}; a root path classification matrix C_{Root}; a classification weighting matrix R, and a document citation matrix D_{ct}. These data matrices are combined such that the resulting document matrix E contains information for each document of its associated ECLA tree.

The classification matrix D_{ec} is a document by ECLA symbol matrix. An entry indicates that a particular ECLA symbol is assigned to a document. The matrix C_{Root} is an ECLA symbol by ECLA symbol matrix. An element in a row is assigned a value of 1 whenever the column corresponds to an ECLA symbol of the root path of the ECLA symbol of the row. Examples of these matrices are shown in Figs. 3 and 4.

Further, a weighting scheme, which is based on the technical specificity of the ECLA symbols as defined in (1), is introduced. This scheme is implemented as a diagonal matrix R, wherein a diagonal element corresponds to an ECLA symbol Θ and an element r_{ii} is calculated by (2) below. The difference as expressed by (2) is applied in order to place emphasis on the leaves of a largest common subtree when calculating the kernel function.

$$r_{ii} = \begin{cases} \sqrt{\text{Specificity}(\Theta)} & \text{if } \Theta \text{ is the root ECLA symbol} \\ \sqrt{\text{Specificity}(\Theta) - \text{Specificity}(\Theta_{parent})} & \text{otherwise} \end{cases} \quad (2)$$

Finally, there is a standard adjacency or citation matrix D_{ct} on the directed citation graph induced by the document's citations. Within D_{ct} the ith row and the ith

$\mathbf{D_{ec}}$	ECLA Root	F	F24	F24H	F24H1/00	F24H1/22	F24H1/40	F24H1/43	F24H8/00	F24H8/00D
Document A	0	0	0	0	0	0	0	0	1	0
B	0	0	0	0	0	0	1	0	1	0
C	0	0	0	0	0	0	0	1	1	0
D	0	0	0	0	0	0	0	1	0	0

Fig. 3 Document classification matrix D_{ec}

$\mathbf{C_{Root}}$	ECLA Root	F	F24	F24H	F24H1/00	F24H1/22	F24H1/40	F24H1/43	F24H8/00	F24H8/00D
⋮	…									
F24H1/43	1	1	1	1	1	1	1	1	0	0
F24H8/00	1	1	1	1	0	0	0	0	1	0
F24H8/00D	1	1	1	1	0	0	0	0	1	1

Fig. 4 Root path classification matrix C_{Root}

column are related to a particular document. An element $d_{ct\,i,j}$ is assigned a value of 1 when document i cites document j.

The calculation of the kernel matrix is carried out by standard matrix operations.

$$E_1 = \operatorname{sgn}(D_{ec} C_{Root})$$

Firstly, the classification matrix and the root path matrix are multiplied and the sign function is applied to each element. This results in a matrix wherein each row corresponds to a particular document and the data in that row contains the information concerning the associated ECLA tree. That is, for any classification symbol present in the classification matrix, the symbols on the root path are added via matrix multiplication.

$$D_2 = D_{ct} + D_{ct}^T$$
$$E_2 = 0.5 \operatorname{sgn}(D_2 D_{ec} C_{Root})$$

Similarly, the citation matrix, the classification matrix and the root path matrix are combined. This yields a matrix which contains information on the ECLA tree of the cited and citing documents. The importance of the citation classification matrix is limited by introducing a weighting value.

$$D_3 = D_{ct} D_{ct} + D_{ct}^T D_{ct}^T$$
$$E_3 = 0.25 \operatorname{sgn}(D_3 D_{ec} C_{Root})$$

There is a further matrix $E3$ for which citations via citations are evaluated.

Combining the matrices E_1, E_2, E_3 and applying the weighting matrix R one obtains:

$$E = (E_1 + E_2 + E_3)R$$

which is used to calculate the **Gram matrix** or **tree kernel matrix**

$$K = EE^T$$

In a final step, the kernel matrix is normalized $K_N : k_{N\,ij} = \frac{k_i k_j}{\sqrt{k_{ii} k_{jj}}}$.

5 Experiment

As a first use case, and to empirically validate that the proposed tree kernel reveals underlying information, a cluster experiment for a set of patent documents was carried out.

The experiment is based on the patent documents assigned the specific ECLA symbol **F24H8/00**. As closely related patents documents are organized in patent families, we used one patent document per patent family for the analysis. There were 588 patent families having the ECLA symbol F24H8/00 assigned thereto. In order to ensure that there was sufficient information available for the clustering process, documents were removed from this set when they could not be related by direct assignments and assignments via citations to at least three ECLA symbols. This resulted in 522 documents for the experiment. The core set was expanded by all the documents which cite a document in the core set as well as all the documents which were cited by the documents in the core set. The bibliographic data of the documents in the expanded set was downloaded from Open Patent Services (OPS). The citation data and the classification data of the documents in the expanded set was then used to calculate the expanded normalized tree kernel matrix, as described in Sect. 4 above using a weighting matrix R in which the values of the low level hierarchy symbols (1st–4th level) was set to 0. A core tree kernel matrix was then formed by removing all rows and columns from the expanded tree kernel matrix which were not associated with a document in the core set. Hence, the core tree kernel matrix contained solely the information about the relateness for any two documents within the core set.

The core tree kernel matrix was passed to the spectral clustering algorithm of Ng et al. (2001). In addition, we tested kernel k-means clustering which also performed reasonably. For comparing the clustering methods as well as for determining the number of clusters, the modularity measure Q according to M. Newman was applied (Newman 2006). Visualization was carried out by kernel principal component analysis (KPCA) (Schölkopf et al. 1996) as implemented by the R package *kernlab* (Karatzoglou et al. 2004). Figure 5 shows the four clusters identified by the clustering experiment. Each solid object represents one of the 522 documents, the

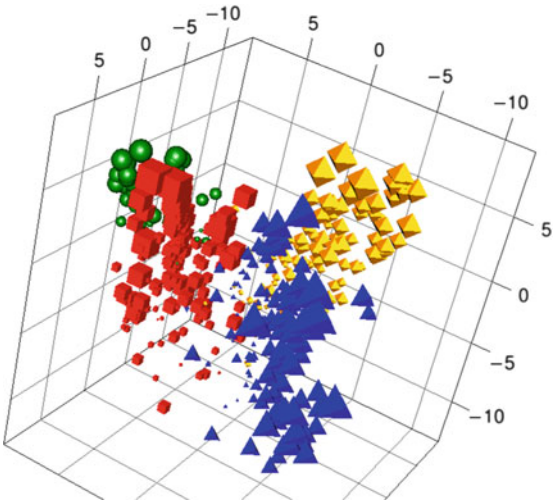


Fig. 5 Visualization of the clustered data set by kernel principal component analysis

color and shape indicating the cluster assignment. The diameter of an object corresponds to the distance of a document to the respective cluster centroid in the feature space. An object has a large diameter when its distance to the centroid is small. Hence, large diameters refer to documents which can be regarded as representative of the cluster. It can be seen in Fig. 5, that within the projection space created by kernel PCA, the largest objects are situated in a corner of a cluster and not in its centre which, however, simply illustrates that the projection space and the feature space do not have the same structure.

The technical concept related to the ECLA symbol F24H8/00 is directed to "Fluid heaters having heat-generating means specially adapted for extracting latent heat from flue gases by means of condensation". When taking a closer look at the technical content of the four clusters, we are able to identify three technical subconcepts which can be labelled "Air heaters", "Water heaters with a water tube heat exchanger" and "Water heaters including fire tubes". In addition, there is a further generic cluster.

6 Conclusions

The presented approach shows how classification and citation information of patent documents can be combined to form a tree kernel on the basis of a strict separation between the definition of the specificity function related to ECLA symbols, the definition of the weighting scheme, the calculation of the relatedness by a kernel function and the final normalizing step. As an example, the kernel function has been applied to cluster patent documents in an effective manner. The results had been

validated by a subject-matter expert. The experiment showed that underlying structures in a specific document set can be revealed. An algorithm-supported application would be a recommendation system to present for a selected document similar documents based on kernel distance measures to users. A suggestion for future work would be to combine the tree kernel with text kernels and direct citation kernels to enhance the results, thereby including further information available in the patent documents.

Disclaimer: This article represents the views and opinions of the authors alone.

We would like to thank the anonymous reviewer for the valuable comments and suggestions.

References

Dickens, D. T. (1994). The ECLA classification system. *World Patent Information, 16*(1), 28–32.

European classification (2008). http://v3.espacenet.com/eclasrch

International Patent Classification (IPC) (2008). http://www.wipo.int/classifications/ipc/en/

Open Patent Services (OPS) (2008). http://www.epo.org/patents/patent-information/free/open-patent-services.html

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software, 11*(9), 1–20.

Li, X., Chen, H., Zhang, Z., & Li, J. (2007). Automatic patent classification using citation network information: An experimental study in nanotechnology. *Proceedings of the 2007 conference on Digital libraries*; JCDL.

Lin, D. (1997). An information-theoretic definition of similarity. *In Proceedings of the 15th International Conference on Machine Learning* (pp. 296–304).

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences USA, 103*, (pp. 8577–8582).

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems* (Vol. 14, pp. 849–856). Cambridge, MA USA: MIT Press.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research, 11*, 95–130.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613–620.

Schölkopf, B., Smola, A., & Müller, K. R. (1996). Nonlinear component analysis as a kernel eigenvalue problem, technical report No. 44, Max-Planck-Institut für biologische Kybernetik.

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press.

A New SNA Centrality Measure Quantifying the Distance to the Nearest Center

Angela Bohn, Stefan Theußl, Ingo Feinerer, Kurt Hornik, Patrick Mair, and Norbert Walchhofer

Abstract In Social Network Analysis (SNA) centrality measures focus on activity (degree), information access (betweenness), distance to all the nodes (closeness), or popularity (pagerank). We introduce a new measure quantifying the distance of nodes to the network center. It is called *weighted distance to nearest center* (*WDNC*) and it is based on edge-weighted closeness (*EWC*), a weighted version of closeness. The *WDNC* will be tested on two e-mail networks of the R community, one of the most important open source programs for statistical computing and graphics. We will find that there is a relationship between the *WDNC* and the formal organization of the R community.

1 Introduction

Until now, SNA centrality measures are based on the idea that a node should be considered more central if it is connected to a lot of other nodes or at least if its friends have many contacts. However, it depends on the question asked to a measure, if this interpretation of centrality makes sense. Imagine a president's wife who is maybe not very interested in politics and who has only a few contacts in a political network, but who has a large influence on her husband. Should she be considered central or not?

The *WDNC* is based on the idea that not only a node's integration into the network is important for its centrality, but also its distance to the center. In the scientific scene, not everyone feels the need to chat with dozens of people every day. However, such people may stay in contact with the network's information brokers, which guarantees him or her access to the most important news. The *WDNC* will be applied to the R (R Development Core Team 2009) mailing lists R-help, designed to discuss users' questions, and R-devel, a communication platform for developers. We will find that the *WDNC* partly reflects the formal organization of the R community.

A. Bohn (✉)
Wirtschaftsuniversität Wien, 1090 Wien, Austria
e-mail: Angela.Bohn@gmail.com

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_63,

2 Methodology

The paper introduces a new measure called *WDNC*. Wasserman and Faust (1997) provide an overview of the most frequently used centrality measures and clustering approaches. The *WDNC* is based on a widely used centrality measure called closeness (Freeman 1979). It is defined as the normalized average distance (length of shortest path) from one vertex to all the others. A modification of closeness, the *EWC* (Bohn et al. 2009), allows to take line values into account. It is defined as

$$EWC(i) = \frac{\sum_j \frac{\mathrm{llv}(i,j)}{d(i,j)}}{\max(\mathrm{lv})(n-1)}, \tag{1}$$

where $\mathrm{llv}(i,j)$ is the (average) last line value on the shortest path between i and j, $d(i,j)$ is the distance between i and j, $\max(\mathrm{lv})$ is the maximum line value in the entire network and n is the network size. The line value between i and j indicates the intensity of interaction and the distance between i and j is the length of the shortest path between them. The shortest path is the minimum number of edges needed to go from i to j. The last line value on the shortest path from i to j is then the line value between k and j, where k is the penultimate node lying on this path. k is identical with i if the distance between i and j is 1. The reason for considering only the last line value instead of using the sum or another aggregation of the line values is a matter of scaling. One could as well use the sum of line values on the shortest i-j-path. In this case, the larger the distance between i and j the more j's contribution to i's *EWC* is influenced by the line values between i and k. Taking only the last line values is more in line with the regular closeness, where each node contributes a certain distance and not a sum of distances. The impression that only the last line value is considered and the others are completely ignored is, however, false. When calculating i's *EWC*, all j are taken into account and thus all the lines lying on shortest paths contribute to i's *EWC*.

Splitting the sum in the enumerator into its summands and marking the distance in which a vertex gains the most *EWC*, corresponds to the definition of the *WDNC*. The *WDNC* of vertex i is defined as

$$WDNC(i) = \left(\inf \operatorname*{argmax}_p \sum_{j \in J_p(i)} \mathrm{llv}(i,j)/d(i,j) \right) - 1 \tag{2}$$

where $J_p(i)$ is the set of all nodes j which can be reached from vertex i by a path of length p. In words: The *WDNC* of a vertex i is the neighborhood p in which it gains the maximum *EWC* minus 1. If the maximum is not unique, infimum chooses the smallest p. The result may be interpreted as a line-weighted distance to the nearest center. The centers are vertices whose *WDNC* is 0. Thus, the *WDNC* combines elements of centrality measures, used to find influential nodes, and community

		p			
		1	2	3	4
i	A	0.136	**0.145**	0.008	0.000
	B	**0.111**	0.111	0.008	0.000
	C	0.111	**0.151**	0.008	0.000
	D	0.049	**0.173**	0.012	0.000
	E	0.062	**0.167**	0.012	0.000
	F	**0.395**	0.019	0.000	0.000
	G	0.136	**0.142**	0.004	0.000
	H	0.012	0.062	**0.095**	0.003
	I	0.012	0.062	**0.095**	0.003
	J	0.012	0.049	**0.074**	0.006

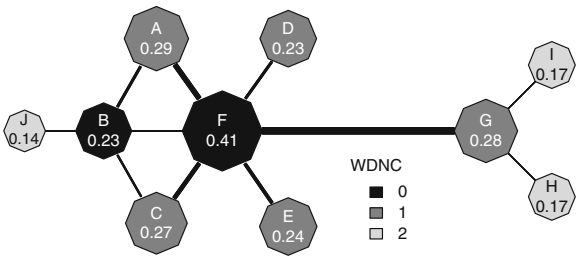


Fig. 1 Example for the calculation of the *WDNC*

detectors, serving to cluster nodes. (The R-code for the *WDNC* can be downloaded from http://r-forge.r-project.org/projects/ewc/.)

Figure 1 shows an example for the calculation of the *WDNC*. The matrix on the left shows the summands of the *EWC* enumerator with bold maxima. The corresponding graph with *WDNC* clusters in gray scale and *EWC* values as labels is on the right. The line strength symbolizes the size of the line values. The black vertices having an *EWC* of 0.23 and 0.41 form the center of the graph. Most of their neighbors' *WDNC* is 1. However, the node having an *EWC* of 0.14 has a *WDNC* of 2, because of its low adjacent line value. This shows that the *WDNC* does not calculate the distance to the nearest center, but the *weighted* distance to the nearest center: Vertices having high line values are closer to the center than nodes having low line values. It is important to notice that the vertices having a *WDNC* of 1 have higher *EWC* values than the black node with an *EWC* of 0.23. This illustrates that the *WDNC* is not the same as calculating *EWC* quantiles.

3 Data and Data Preparation

The characteristics of the *WDNC* are investigated using network data of the R-help and R-devel mailing lists during 2008. They serve to discuss questions from R developers and users, therefore they contain interesting information about a part of their social structure. Every e-mail sent to the mailing list is forwarded to all subscribers. They can be downloaded as compressed text files from https://stat.ethz.ch/pipermail/r-devel/ and https://stat.ethz.ch/pipermail/r-help/, respectively.

Transforming Thread Trees to a Social Network

Usually, mailing lists are represented as thread trees showing the referencing links between e-mails. Each e-mail has a message-ID and follow-ups additionally have reply-to IDs allowing to build thread trees (Feinerer et al. 2008). The next data preparation step consisted in transforming the thread trees, where nodes represent

e-mails, in such a way that nodes represent e-mail authors. We drew an edge between an author and his or her "forefathers" in the thread tree. The networks are represented as weighted matrices, where the weights correspond to the number of e-mails exchanged between two authors. To calculate the *WDNC*, we need the networks to be strongly connected. As the largest strongly connected subgraph (component) cover only a small part of the network members we symmetrized the networks using the sum of incoming and outgoing arc weights (sum of sent e-mails and received e-mails) and only the largest component was considered. The other components (42 in R-help and 15 in R-devel) have only one to three members and are therefore negligible.

Finding Aliases

The second data preparation step consisted in finding aliases, as authors may have several different user names and e-mail addresses. Like Bird et al. (2006) we first normalized the user names and e-mail addresses, then we used the Levenshtein distance (Levenshtein 1966) to find clusters of similar names. To increase the probability of finding all aliases, we allowed a distance of 0.3 between the names within one cluster. Thus, each cluster contained a number of strings that differed in at most 3/10 of the symbols. We checked those clusters manually and rejected 60% of them, so we expect to have found most aliases. This way, the R-help network was reduced from 5,128 to 4,065 nodes and the R-devel network from 837 to 652.

Description of R-Help Network

The largest component of the network has 3,672 nodes, its diameter (length of longest shortest path) is seven, the average degree (number of direct neighbors) is 11.8 and the median degree is 4. Each network member wrote 7.6 e-mails on average. The maximum of e-mails sent was 1,071 by Brian Ripley. About 1,640 people wrote only one e-mail. About 76% of the line values (number of e-mails exchanged between two authors) is 1. The maximum of e-mails exchanged between two authors was 72 (Gabor Grothendieck and Brian Ripley) and their mean is 1.5.

Description of R-Devel Network

The largest component of the R-devel network has 566 nodes. Its diameter is 6. As the network is much smaller, the average degree is 8.5, but the median degree is also 4. Brian Ripley is by far the most active author in the R-devel network. He sent 522 e-mails and his degree is 332. The second most active author, Duncan Murdoch, wrote only 255 e-mails and his degree is 177. Most line values (67%) are 1 and the their mean is 1.9. The maximum of e-mails exchanged between two authors was 63 (Brian Ripley and Duncan Murdoch).

4 Results

In this section, we will apply the *WDNC* presented in Sect. 2 to the networks described in Sect. 3. The results will be compared to other centrality measures. Finally, the informal structure of software development will be compared to the formal organization.

4.1 R-Devel Network

In the R-devel network, we identified five very central authors using the *WDNC*: Peter Dalgaard, Gabor Grothendieck (GG), Martin Mächler, Duncan Murdoch, and Brian Ripley (BR). They are in close contact to each other, so the network does not to have several separated centers with each having its own community, but it is monocentric. BR is the most active author in terms of degree and number of e-mails sent. Many vertices adjacent to BR do not have any other contacts. He prevents the network from being split into many small components. In contrast, some of BR's neighbors are more active and they are well connected to other network members, so they may be considered to be the core of the network. 69% of the network members have a *WDNC* of 1. Most of them are neighbors of the central cluster. 28% have a *WDNC* of 2 and most of them are neighbors of those having a *WDNC* of 1.

4.2 R-Help Network

In the R-help network we identified three very central authors: GG, Jim Holtman, and BR. Like in the R-devel network, the central nodes are in close contact to each other. However, in this network, each of the vertices having a WDNC of 0 have a community that is partly separated from the others. Furthermore, all nodes in the central cluster are comparably active. These observations indicate, that BR's position is not as marked as in the R-devel network.

4.3 Empirical Evidence of the Usefulness of the WDNC

As the *WDNC* defines a vertex' importance according to its weighted distance to the nearest center, it is crucial to know whether the choice of centers is reasonable. Figure 2 shows boxplots of centrality measures for each *WDNC* cluster (x-axis) in the R-help network. The vertices having a *WDNC* of 0 are far more central than the other clusters according to degree and pagerank (Brin and Page 1998). Compared to closeness and *EWC*, the difference between cluster 0 and the others is smaller, because the *WDNC* is based on these measures. The corresponding boxplots of the R-devel network are very similar (Fig. 3), so we conclude that in the mailing list

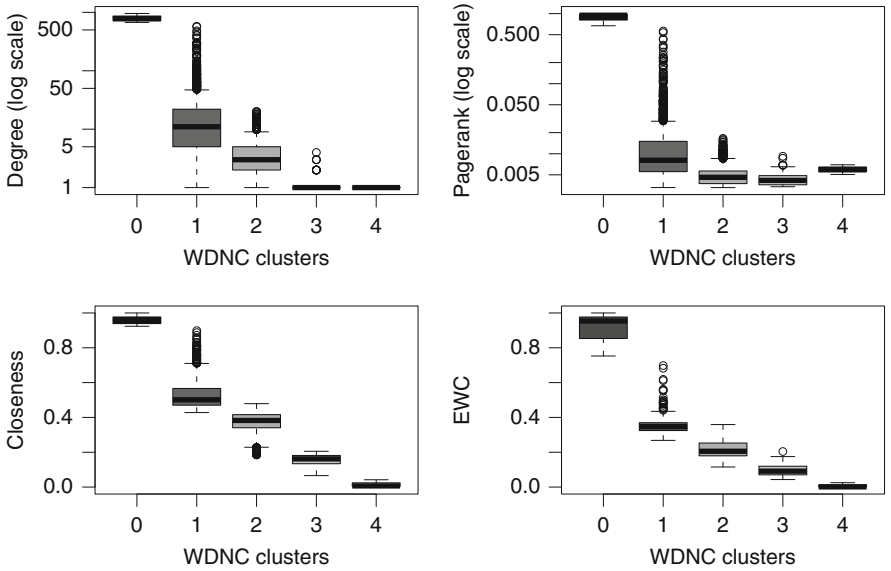


Fig. 2 Boxplots of *WDNC* vs. centrality measures in the R-help network

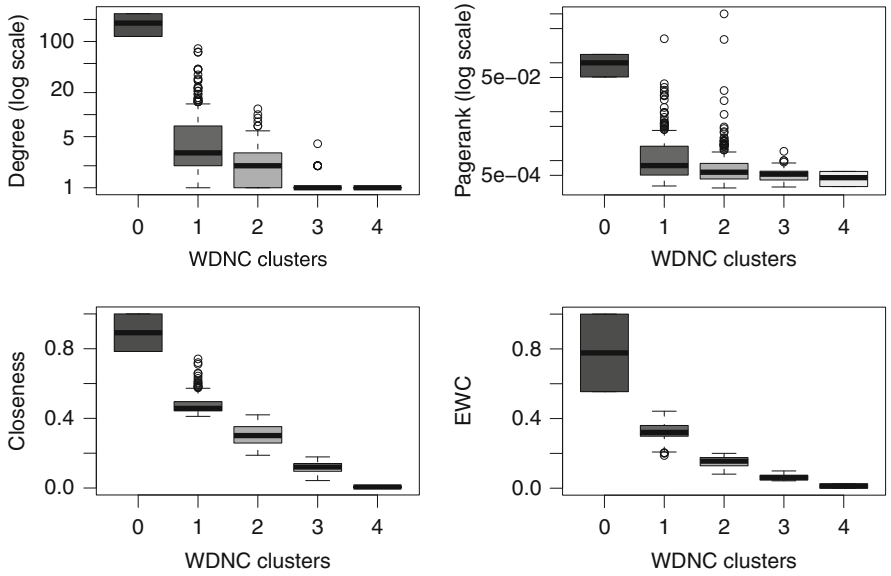


Fig. 3 Boxplots of *WDNC* vs. centrality measures in the R-devel network

networks, the algorithm chose a few very central vertices to have a *WDNC* of 0. Nodes with a *WDNC* of 1 are clearly less central, however, the amount of outliers in this cluster (139–140 in R-help and 23–25 in R-devel) indicates that it is heterogeneous.

Table 1 Cross-classified table of *WDNC* vs. developer and user groups

R-devel						R-help					
WDNC	c.d.[a]	m.d.[b]	o.d.[c]	users	sum	*WDNC*	c.d.[a]	m.d.[b]	o.d.[c]	users	sum
0	4	0	1	0	5	0	1	0	1	1	3
1	11	11	186	185	393	1	11	13	411	961	1,396
2	0	2	64	94	160	2	3	3	581	1,552	2,139
3	0	0	3	5	8	3	0	2	33	96	131
4	0	0	0	0	0	4	0	0	1	2	3
sum	15	13	254	284	566	sum	15	13	1,027	2,612	3,672
mean	**0.7**	**1.2**	**1.3**	**1.4**	1.3	mean	**1.1**	**1.4**	**1.6**	**1.7**	1.7

[a] Core developers
[b] Main developers
[c] Other developers

4.4 WDNC Compared to Formal R Organization

The formal organization of R is not as strictly defined as in software companies. However, the R community can be roughly divided into several groups of developers and users. There are 19 core developers and 44 main developers who are mentioned on the R Core Team website (http://www.r-project.org/contributors.html). In addition, there are hundreds of other developers whose names can be obtained from the R package descriptions on CRAN (http://cran.r-project.org/). The group of users can be divided into active and passive users. Active users report bugs, make suggestions for improvements and write to the mailing lists. Passive users do not communicate their experiences, but only use the software. Thus, the mailing lists contain only information about active users. However, we cannot distinguish between the different kinds of active users. Table 1 shows a cross-classified table of *WDNC* vs. developer and user groups.

It shows that, although R-devel is intended for developers and R-help for users, a separation between users and developers is only partly realized: Half of the R-devel authors are users and 29% of the R-help authors are developers. (Note that some developers might be classified as users if their package is not yet on CRAN.) However, inside the networks, the behavior of the groups differs. If we take the membership of an author to a certain developer or user group as an indicator for the level of commitment of this author to R, where the membership to the core developers corresponds to highest commitment and the membership to the user group means lowest commitment, we see that the mailing list behavior reflects these differences: The core developers have the lowest average *WDNC* in both networks (0.7 and 1.1), which means that they are most central. The group of main developers is slightly less central (1.2 and 1.4) and the other developers have an average *WDNC* of 1.3 and 1.6. Finally, many users are located at the periphery, which results in an average *WDNC* of 1.4 and 1.7. (Like any other centrality measure, the *WDNC* of a vertex can only be interpreted in comparison to nodes of the same network and not across networks: An average *WDNC* of 1.4 can indicate a central position in one

network and a peripheral position in another.) Thus, a low *WDNC* is associated with high commitment.

5 Conclusion and Discussion

This paper introduced a combination of centrality measure and clustering approach called *WDNC*. The *WDNC* was applied to the OSS mailing lists R-devel and R-help. We found that the network structure of both mailing lists are similar: They are monocentric and dominated by a few very active e-mail authors staying in close contact to each other. This can be explained by the fact that the mailing lists do not reflect a stringent separation between developers and users. However, the *WDNC* reveals that the behavior of users and types of developers differs. If we take a developer's formal role as indicator for his or her commitment to R, where the membership to the core development group indicates highest commitment and being a user indicates lowest commitment, we see that a low *WDNC* is associated with high commitment. Thus, the level of commitment tends to be reflected by a central and influential position in the mailing lists. However, the validity of the results is restricted to the communication via mailing lists which capture only a small part of the social behavior. Although the data structure did not allow to use the directed version of *WDNC*, it can be useful in other applications, for example to distinguish question-people from answer-people.

References

Bird, C., Gourley, A., Devanbu, P., Gertz, M., & Swaminathan, A. (2006). Mining email social networks. In *Proceedings of the 2006 international workshop on Mining software repositories, Shanghai, China* (pp. 137–143).

Bohn, A., Walchhofer, N., Mair, P., & Hornik, K. (2009). Social network analysis of weighted telecommunications graphs. Report 84, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Research Report Series.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1–7), 107–117.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software, 25*(5), 1–54.

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks, 1*, 215–239.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Technical Report 8, Soviet Physics Doklady.

R Development Core Team. (2009). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Wasserman, S., & Faust, K. (1997). *Social network analysis – methods and applications*. Cambridge: Cambridge University Press.

Mining Innovative Ideas to Support New Product Research and Development

Dirk Thorleuchter, Dirk Van den Poel, and Anita Prinzie

Abstract Here, we present an approach for automatically identifying the innovative potential of new technological ideas extracted from textual information. The starting point of each innovation is a good and new idea. Unfortunately, a high percentage of innovations fail, which means many ideas do not have the potential to become an innovation in future. The innovation process from a new idea as starting point via research, development, and production activities through to an innovative product is very cost- and time-consuming. Therefore, the aim of our work is to identify the innovative potential of new technological ideas to improve the performance of the innovation process.

We extract new technological ideas from provided textual information. We also identify innovative technology fields by analysing relationships among technologies. All identified ideas are assigned to innovative technology fields by using text mining and text classification methods. Technological ideas in these fields are presented to the user as innovative ideas and might be used as starting point for new product research and development divisions.

1 Introduction

The word innovation refers to the latin terms novus (that means new) and innovation (that means something is newly created). An innovation includes a new idea (Guiltinan and Paul 1991) as well as its realization e.g. as innovative product that is successful in marked. Therefore in economical sense, we talk about innovations if the newly created object increases producer or customer value (Mckeown 2008).

To create an innovation, an innovation process can be used. It has the aim to lead a new idea to an innovative product. Therefore, the starting point of the innovation process is a new technological idea (Möslein and Matthaei 2008). Based on this idea,

D. Thorleuchter (✉)
Fraunhofer INT, D-53879 Euskirchen, Appelsgarten 2, Germany
e-mail: Dirk.Thorleuchter@int.fraunhofer.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_64,

a research process starts. The result is probably a prototype that is developed further in a developing process. After this developing process a production process starts and leads to a product (Bürgel et al. 1996). If this product is successful in market that means it increases producer or customer value then it is an innovative product and the idea standing behind this innovation can be defined as innovative idea. However, by use of this economical definition, we only can identify innovative ideas subsequent to the innovation process that means after they become successful products in marked.

Unfortunately, the innovation process is very cost- and time-consuming (Disselkamp 2005) and a high percentage of innovations fail. Therefore, the aim of our work is to identify the innovative potential of new technological ideas before selecting them as starting ideas. This probably can improve the performance of the innovation process.

2 Rationale Behind Mining Innovative Ideas

Our definition of a technological innovation is based on bibliometrical analysis as described in Reiß (2006). There, it is shown that innovations normally do not occur alone but together with several further innovations. These groups of innovations are based on a common innovation field. Innovation fields are newly appeared technologies or scientific disciplines that occur on the borders of established technologies or scientific disciplines. This means they occur between at least two technologies or scientific disciplines that are not related. A definition of possible relationships is given in Sect. 6. Therefore, innovations can be classified as interdisciplinary products. The (innovative) ideas behind these innovations also are of an interdisciplinary nature and they also occur together in an innovation field.

Our idea definition derived from technique philosophy (Rohpohl 1996). There, a technological idea consists of two things: a means and an appertaining purpose (Thorleuchter et al. 2010). Therefore, we define an idea as a text phrase. This text phrase consists of domain specific terms that occur together in textual information. These terms can be divided up into two subsets. The first subset should represent a means and the second subset should represent a purpose. An example for an idea is a nanomagnet (the means) that can be used to switch electronic signals (the appertaining purpose). This definition is used to identify interdisciplinary ideas by assigning means and purpose of an idea to different non-related, established technologies or scientific disciplines.

To classify ideas as innovative, we have to identify several interdisciplinary ideas that occur together in an innovation field. For this, we firstly have to provide technological context information containing descriptions of established technologies or scientific disciplines and we have to define their relations.

Secondly, we have to classify ideas as interdisciplinary by assigning means and purposes to established technologies or scientific disciplines that are not related. For example, if a means from a bionic idea can be assigned to biology and the

appertaining purpose can be assigned to technological engineering then the bionic idea is interdisciplinary. This gives a hint that the combination of biology and technological engineering is probably an innovation field.

To be sure that it is really an innovation field, we thirdly have to find several further interdisciplinary ideas that can be assigned to the same non-related technologies or scientific disciplines combination and classify all the interdisciplinary ideas in this field as innovative ideas.

3 Process of Mining Innovative Ideas

This approach uses an existing idea mining approach (Thorleuchter 2008) that supports users to identify means and purposes in text phrases (see Sect. 4). Then, we provide descriptions of scientific categories as context information (see Sect. 5). Both the means and purpose of extracted new and useful ideas are assigned to several scientific categories by use of multi-label classification (see Sect. 7). After this, we compare each scientific category from means to each scientific category from purpose to find out relationships between them (see Sect. 6). Figure 1 shows an example for the processing of this approach.

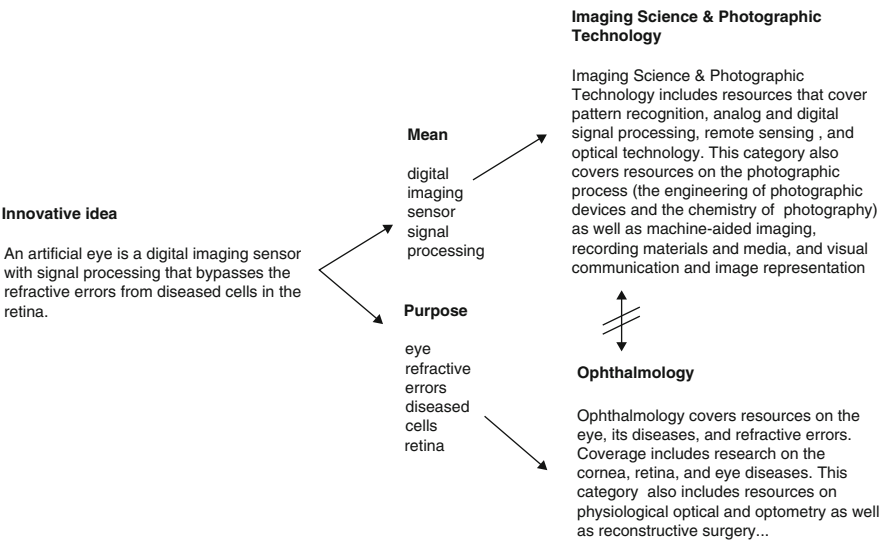


Fig. 1 Means and purpose are extracted from an idea and are assigned to different scientific categories. If the categories that are assigned by the means are not related to categories that are assigned by the purpose then the idea is interdisciplinary. If several ideas also are interdisciplinary concerning these categories then the combination of both categories is defined as innovation field and ideas from this field are presented as innovative ideas

If a means is assigned to several scientific categories and the appertaining purpose is assigned to further scientific categories that are not related to any scientific category of the means then the corresponding idea is interdisciplinary. If several further ideas also are interdisciplinary concerning at least two of the above mentioned scientific categories then we define the combination of these scientific categories as innovation field. For this, the user provides the smallest number of interdisciplinary ideas that is sufficient to define such an innovation field. Ideas from these innovation fields are classified as innovative ideas.

4 Acquisition of Ideas

Our approach is based on technological ideas. The user extracts them from provided textual information e.g. patent data. He is supported by a further approach that automatically extracts new and useful ideas from textual information as presented in Thorleuchter (2008). This approach extracts textual phrases that represent new and useful ideas. Additionally for each idea, it identifies terms that represent a means as well as terms that represent the appertaining purpose. This is used as input for our approach.

5 Acquisition of Technological Context Information

To provide technological context information, we focus on scientific categories. We can find an overview of current scientific categories in the science citation index (SCI). This index is built on bibliographic information, author abstracts, and cited references from about 3,700 science and technical journals. The content of these highly cited journals is assigned to 172 scientific categories. The official description of all categories in the SCI is available in scope notes (Institute for Scientific Information 1997) that is manually created, of good quality, and up to date. We use this description as technological context information for our approach.

6 Relationship among Scientific Categories

After providing descriptions of scientific categories that represent technological context information, the next step is to identify relationships among these scientific categories.

In general, we identify two different kinds of relationships (Geschka et al. 2005). One kind of relationship is that technologies can be similar to other technologies. They deal with the same technology field but have a different focus. The descriptions

of two similar technologies also are similar because they both contain the same domain specific terms by describing the technological field.

A further kind of relationship is that technologies are related in a substitutive, integrative, predecessor or successor way. If technologies are related in this way then they deal with the same application field. Their descriptions also are similar because they both contain the same domain specific terms representing the application field.

The descriptions of the scientific categories in scope notes contain terms representing the technological field as well as terms representing potential application fields. If we identify similar terms in descriptions of two different scientific categories then both categories are related according to at least one kind of relationship. Therefore, related categories are identified by comparing category descriptions among each other.

Comparing is done by transforming category description to term vectors in vector space model. For this, terms in the descriptions are tokenized (Feldman and Sanger 2007) by using the term unit as word, stop word filtered by using a standard stop word list (Lustig 1986), and stemmed (Hotho et al. 2005) using a dictionary-based stemmer combined with a set of production rules (Porter 2008) to give each term a correct stem. The production rules are used when a term is unrecognizable in the dictionary. Vectors representing scientific categories can be compared using similarity measures in combination with the fuzzy alpha cut method (Abebe 2000) and two categories are classified as related if the corresponding similarity measure result value is greater than or equal to alpha. For comparing, we prefer the well-known Jaccard's coefficient measure (Ferber 2003) because it considers well the different sizes of both vectors.

7 Classification of Ideas

Each selected idea consists of a set of terms that represents a means and of a set of terms that represents an appertaining purpose. To identify an interdisciplinary technological idea we have to assign both sets to scientific categories. Both sets of terms are stop word filtered and stemmed as described in Sect. 6. For multi-label classification, we transform these sets to term vectors in vector space model and compare them with term vectors from each scientific category. For comparing, we also use Jaccard's coefficient measure in combination with the fuzzy alpha cut method. As a result, means and purposes are assigned to scientific categories only if the appertaining Jaccard's coefficient result value is greater than or equal to alpha.

Each means and each purpose of a new idea is probably assigned to several scientific categories. To identify relations, we compare each scientific category from means to every single scientific category from purpose as described in Sect. 6. If we cannot find any relationships then the new idea is interdisciplinary and each of these scientific category combinations from means and purpose is probably an innovation field. If we identify at least n interdisciplinary ideas that can be assigned to one specific scientific category combination then we define an innovation field on this basis.

The user provides the smallest number n of interdisciplinary ideas that are sufficient to define such an innovation field.

8 Results and Evaluation

We present a heuristic approach for automatically identifying the innovative potential of new technological ideas. The extraction of ideas and the identification of terms that represent means and purposes is already evaluated in Thorleuchter (2008). Therefore, the evaluation is limited to the further steps of our approach. The evaluation of new ideas for their innovative potential based on current context information. For this, scientific categories in the science citation index as current technological information described in scope notes (Institute for Scientific Information 1997) are used.

The approach extracts 1,000 new ideas from randomly selected patents because patent descriptions consist of new ideas that also are innovative. However, not all new ideas are innovative in terms of the technological innovation definition in Sect. 2. Five hundred ideas are used as training examples to obtain the optimal parameter values and 500 ideas are used as test set to validate and compare the model. To evaluate the results of the approach, we use precision and recall measures commonly used in information retrieval based on true positives, false positives, and false negatives. For this, the ground truth for our evaluation is defined. Therefore, a human expert classifies the 1,000 new ideas as innovative or as non-innovative.

The approach depends on three parameters (n, α_1, α_2). The smallest number (n) of interdisciplinary ideas that is sufficient to define an innovation field gives a hint concerning the innovative potential of the new idea. If the number n is large then we only obtain ideas as result items that probably consist of a very high innovative potential. This is because we identify many ideas that are classified concerning a specific non-related combination of scientific categories. Here, we have a high probability that this category combination represents an innovation field. If the number n is small e.g. it equals one then we get all interdisciplinary ideas as result items regardless weather they consists of high or low innovative potential. This is because every idea – that is classified concerning a specific non-related combination of scientific categories – is presented as innovative idea. We estimate that an optimal value of n is between $4 \leq n \leq 8$.

After this, the alpha cut of Jaccard's coefficient results are estimated. The first alpha cut is the set of all terms that represents a means or a purpose such that the corresponding result value by comparing this set to a scientific category is greater than or equal to α_1. With the second alpha cut we identify two related scientific categories only if the appertaining Jaccard's coefficient result value is greater than or equal to α_2. If α_1 is too small or too large then means and purposes are not classified correctly. If α_2 is too small or too large then the identification of relationships among scientific categories fails. This leads both to a small precision and to a small recall value. An optimal value of α_1 and α_2 is estimated between $5\% \leq \alpha_1, \alpha_2 \leq 20\%$.

To investigate the dependency of the approach on the parameters, we explicitly check if the parameter values are identifiable on the training set. These values are used to compute precision and recall on the test set. For this, we use the estimations for $n \in \{4, 5, \ldots, 8\}$ and the percentages $\alpha_1 \in \{5\%, 6\%, \ldots, 20\%\}$ and $\alpha_2 \in \{5\%, 6\%, \ldots, 20\%\}$. We identify $5 \cdot 16 \cdot 16 = 1280$ different parameter combinations of (n, α_1, α_2). The training set is used to compute average precision and recall for each parameter combination to identify the optimal parameter values with a maximal F-measure. The F-measure is used because precision and recall are equally important. As a result, parameter values $n = 5$, $\alpha_1 = 14\%$, and $\alpha_2 = 16\%$ are identified. These parameter values are used to compute precision and recall for each test example and the average precision and recall values for all test examples. We get a precision value of 38% and a recall value of 30%. A precision value of 38% means that if this approach predicts 100 ideas as innovative ideas then 38 of them are innovative. A recall value of 30% means that if there are 10 innovative ideas in the provided text then this approach identifies three of them.

We compare this approach to a baseline model because we are not aware of other approaches for identifying the innovative potential of ideas at the present time. A positive class probability of 5% is already calculated by human experts. This leads to a 5% precision at 30% recall for a random prediction and it shows that this approach is much better than random. We think that the results are sufficient to proof the feasibility of our approach.

Using the 500 new ideas from the test set, the approach automatically computes several innovation fields. We present examples for these innovation fields. They can be found between 'Health Care Sciences and Services' and 'Computer Science, Artificial Intelligence' (e.g. the use of methods from artificial intelligence for health care applications), between 'Imaging Science and Photographic Technology' and 'Medical Informatics', between 'Remote Sensing' and 'Tropical Medicine', and between 'Computer Science, Theory and Methods' and 'Psychiatry'. Then, the approach identifies ideas from these innovation fields as innovative ideas.

This approach can be re-evaluated by using our application for mining innovative ideas (see http://www.text-mining.info). There, the web based application that is programmed in perl/ruby and all texts that are used for evaluation are presented. The application extracts ideas from a provided text, creates terms representing means and purposes, identifies innovation fields, and classifies the ideas as (non-) innovative ideas.

9 Outlook

This work shows that the automatic identification of the innovative potential of new technological ideas is feasible using text classification and specific technological definitions. Further work should aim at enlarging and optimizing this approach e.g. by identifying further properties of innovative ideas. A second avenue of further research could take the granularity of the context information into account e.g. by

using technologies rather than scientific categories. This also probably leads to an increasing precision and recall.

References

Abebe, A. J., Guinot, V., & Solomatine, D. P. (2000). Fuzzy alpha-cut vs. Monte Carlo techniques in assessing uncertainty in model parameters. In *Proceedings of the 4-th International Conference on Hydroinformatics*. Iowa City, USA.

Bürgel, H. D., Haller, C., & Binder, M. (1996). *F&E-Management* (p. 85). Vahlen: München, 1996.

Disselkamp, M. (2005). *Innovationsmanagement: Instrumente und Methoden zur Umsetzung im Unternehmen* (p. 179). Wiesbaden: Gabler Verlag.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data* (p. 318). Cambridge: Cambridge University Press.

Ferber, R. (2003). *Information retrieval* (p. 78). Heidelberg: dpunkt.verlag.

Geschka, H., Schauffele, J., & Zimmer, C. (2005). Explorative technologie-roadmaps - eine methodik zur erkundung technologischer entwicklugslinien und potenziale. In M. G. Möhrle and R. Isenmann (Eds.), *Technologie-roadmapping* (p. 165). Berlin, Heidelberg: Springer.

Guiltinan, J. P., & Paul, G. W. (1991). *Marketing management: Strategies and programs* (p. 196). NY, U.S.A.: McGraw-Hill.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. *LDV Forum, 20*(1), 19–26.

Institute for Scientific Information (Ed.) (1997). *SCI Journal Citation Reports*. Philadelphia, PA: Institute for Scientific Information.

Lustig, G. (1986). *Automatische Indexierung zwischen Forschung und Anwendung* (p. 92). Hildesheim: Georg Olms Verlag.

Mckeown, M. (2008). *The truth about innovation*. Harlow: Pearson Education.

Möslein, K. M., & Matthaei, E. E. (2008). *Strategies for innovators: A case book of the HHL open school initiative* (p. 13). Wiesbaden: Gabler Verlag.

Porter, M. F. (2008). An algorithm for suffix stripping. *Program, 14*(3), 130–137.

Reiß, T. (2006). Innovationssysteme im Wandel – Herausforderungen für die Innovationspolitik. In B. Müller and U. Glutsch (Eds.), *Fraunhofer-Institut für System – und Innovationsforschung – Jahresbericht 2006* (p. 10). Fraunhofer ISI edition.

Rohpohl, G. (1996). Das Ende der Natur. In L. Schäfer and E. Sträker (Eds.), *Naturauffassungen in Philosophie, Wissenschaft und Technik* (Bd. 4, pp. 143–163). Freiburg, München: Alber.

Thorleuchter, D. (2008). Finding technological ideas and inventions with text mining and technique philosophy. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme and R. Decker (Eds.), *Data analysis, machine learning, and applications* (pp. 413–420). Berlin, Heidelberg: Springer.

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010). Mining Ideas from Textual Information. *Expert Systems With Applications*, doi:10.1016/j.eswa.2010.04.013.

The Basis of Credit Scoring: On the Definition of Credit Default Events

Alexandra Schwarz and Gerhard Arminger

Abstract The concrete circumstances which lead to the definition of a certain credit as default or non-default are usually not discussed in the literature. Since we have access to a large data set of complete payment histories of relatively short-termed installment credits, we can investigate a possible solution to the problem of defining and detecting defaulted credits. We propose a definition of credit default events where we adopt the patterns of payment, a common measure for the control of accounts receivable, to the case of installment credits. In addition, we define indicators of individual payment performance which can be used for monitoring payment behavior on-line, once a credit scoring system has been implemented in practice. Consequently, the update of a scoring model need not depend on the information of closed credits, which becomes only available at the end of the payment term at the earliest.

1 Introduction

Credit scoring is the standardized process of analyzing and classifying the payment behavior of individuals, i.e. private consumers or companies. This process is based on statistical methods for estimating the individual propensity to show the expected payment behavior, e.g. to settle payment by installments regularly. On the basis of this assessment, credit applicants are assigned to one of two (or even more) classes. These classes represent sufficient and insufficient payment, or creditworthy and not creditworthy customers, respectively.

The statistical techniques that cover the process of modeling individual credit risks are widely discussed in the literature (e.g. by Hand and Henley 1997; Thomas et al. 2002). In contrast, we find only few statements on how the dependent variable

A. Schwarz (✉)
Schumpeter School of Business and Economics, University of Wuppertal,
Gaußstr. 20, 42097 Wuppertal, Germany
e-mail: schwarz@statistik.uni-wuppertal.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_65,

default yes/no in a scoring model is defined. Even in statistical publications, this definition is always said to be given, but not described. Nonetheless, defining credit default events is a critical task within the process of modeling credit risk as any such definition is needed to operationalize the key dependent variable, e.g. creditworthiness. It can be assumed that this lack of information is due to confidentiality reasons because the definition of credit default gives direct insight into a bank's or company's internal calculations, its credit policy and marketing strategy.

In Sect. 2 we briefly describe the empirical data set which we analyze in the course of this paper. To arrive at a possible definition of credit default we adopt the patterns of payment, a common measure for the control of accounts receivable, to the case of installment credits in Sect. 3. Based on a measure of profitability that can be derived from the payment patterns we classify events of default and non-default. In Sect. 4 we define indicators of individual payment performance for the monitoring of payment behavior and evaluate them with respect to their potential to detect defaults on-line, i.e. already during the payment process. The paper closes with a discussion in Sect. 5.

2 Data Set of Individual Payment Histories

To exemplify our proposals for defining and monitoring credit default events, we analyze 33,986 installment purchases of household appliances, each of them paid off by 15 regular installments. For the company granting these credits, 13.29% of the financed amounts remain uncollectible. The data set consists of all credits for which the due date of the first rate of payment lays between March 01, 2004 and August 31, 2004. Therefore, we analyze a complete cohort of credits. For each credit, we observe the payment history running from the due month of the first installment until March, 2007. The payment histories are given in the form of monthly account balances. This implies an exact installment plan with a due date of each installment, but we are not given the exact date at which payments are made. Consequently, we analyze the data related to each installment credit on a monthly basis.

3 A Payment-Pattern Approach to the Identification of Credit Default Events

A loan contract in the special form of a payment in installment always involves an installment plan which documents the due dates and due amounts of repayment. These due, expected payments can be compared to the actual payments of a debtor by means of the individual account balances. The basic idea of the proposed classification is to balance expected and actual payments of debtors at every point in time at which payments are expected. By evaluating this pattern of payments and the profitability of the involved accounts we can determine the maximum period of deficient payment which is acceptable for financial purposes.

3.1 The Patterns of Payment

In the context of sales on credit the receivable balance pattern "is the proportion of any month's sales that remains outstanding at the end of each subsequent month" (Johnson and Kallberg 1986). This proportion is expected to decay over the subsequent months. Therefore, it is tracked by simply following the percentages over time. The collection pattern is the mirror image of the receivable balance pattern, giving the cumulative collections of the subsequent months in percent of credit sales. In the following, we define suitable patterns of payment for the case of installment credits. A description of the original procedures is given in Stone (1976).

Suppose we observe the complete payment history of n installment credits $i = 1, \ldots, n$ with total financed amounts y_i. We also suppose that these credits are paid off by an equal number T of installments, and that payments are due at regular intervals. Hence, payments are observed at points in time $t = 1, \ldots, T, \ldots, T + h, \ldots, T + H$ with every t denoting an observation point of due installments. Then T denotes the total number of installments and at the same time the end of the agreed payment term, and H is the number of points in time $h = 1, \ldots, H$ at which we observe payments after the end of the regular payment term. Hence, $T + H$ describes the end of our complete observation period.

Let $y_{i,t}$ denote the due amount of payment of credit i at time t, which is the agreed installment at time t, and let $x_{i,t}$ denote the respective amount actually paid at time t. Then

$$X_k = \sum_{t=1}^{k} x_t \quad \text{with} \quad x_t = \sum_{i=1}^{n} x_{i,t} \tag{1}$$

are the cumulated payments actually made until k with $k \in \{1, \ldots, T + H\}$. Equivalently, the cumulated expected payments until k are denoted by

$$Y_k = \sum_{t=1}^{k} y_t \quad \text{with} \quad y_t = \sum_{i=1}^{n} y_{i,t} \tag{2}$$

Consequently,

$$\Delta_k = \sum_{t=1}^{k} \delta_t \quad \text{with} \quad \delta_t = \sum_{i=1}^{n} (y_{i,t} - x_{i,t}) \tag{3}$$

are the cumulated outstanding payments at time k. Obviously, $Y_k = X_k + \Delta_k$ at each k. In this retrospective analysis of payments the collection pattern over $T + H$ points of observation can be calculated as the respective cumulated payments in % of the overall financed amount $Y = \sum_{t=1}^{T} y_t$, i.e. X_k / Y for all k. Respectively, the receivable balance pattern is given by Δ_k / Y, that are the respective cumulated outstanding amounts in % of the total expected payment.

3.2 Measurement of Profitability

Our definition of credit default events is based on the profitability of accounts which can be measured using the patterns of payment described above. To illustrate this approach we assume that an account is (still) profitable at time k if the additional costs caused by deficient payment that occurred up to k are strictly smaller than the absolute profit margin generated by payments made up to k. For measuring the profitability P_k of accounts at time k, we introduce a weight a for the cumulated payments and a weight c for the cumulated outstanding amounts Δ_k. Here a (in %) may be interpreted as the profit margin and c may denote the rate of additional costs (in %), especially interest charges, involved in the cumulated outstanding amounts. Then the profitability can be measured as

$$P_k = a \cdot X_k - \sum_{t=1}^{k} c \cdot \Delta_k \tag{4}$$

Here the cumulated additional costs $\sum_{t=1}^{k} c \cdot \Delta_k$ incorporate the so-called revolving effect of credit which occurs if deficient payments are protracted over a certain period. Let t^* denote the minimum k of all observation points for which $P_k \leq 0$:

$$t^* = \min_{k=1,\ldots,T+H} (k \mid P_k \leq 0) \tag{5}$$

This means, t^* is the point in time of the period of deficient payments at which the performance of credits is no longer acceptable, whereas $t^* - 1$ denotes the last point in time of the period of acceptable performance. This leads to the following classification rule concerning the definition of the default event Z: Credit i is assigned to the class of bad accounts if it contributes to the overall loss, that is, it shows an outstanding amount at t^*. Otherwise credit i is assigned to the class of good accounts. With

$$\delta_{i,t^*} = \sum_{t=1}^{t^*} (y_{i,t} - x_{i,t}) \tag{6}$$

denoting the sum of outstanding amounts for credit i at t^* the classification rule is

$$z_i = \begin{cases} 1 & \text{if } \delta_{i,t^*} > 0 \\ 0 & \text{else} \end{cases} \tag{7}$$

From a financial perspective, classifying credits as defaults based on t^* as defined in (5) means that the costs of financing the outstanding amounts exceed the profit made by the received payments. Hence, it may be useful to generalize the right hand side of (5) by $P_k \leq \tau$ where $\tau = 0$ in the above example.

3.3 Application to the Empirical Data Set

In the empirical example we observe the individual payment histories of $n = 33{,}986$ credits with an agreed number of installments $T = 15$ and an additional observation period of $H = 17$ months where we expect accounts to be finally balanced. Therefore, the complete observation period is $T + H = 32$ months. Figure 1 shows the patterns of payment in % of the overall financed amount for each t, that are the expected pattern (Y_k/Y), the collection pattern (X_k/Y) and the receivable balance pattern (Δ_k/Y).

To exemplify the measurement of profitability we choose $a = 5\%$, $c = 1\%$ for the regular payment term $t = 1, \ldots, 15$ and $a = 5\%$, $c = 2\%$ for the additional period $t = 16, \ldots, 32$ after the end of the agreed payment term. In Fig. 1 the profitability is given in % of the expected profit $(P_k/(a \cdot Y_k))$. The left vertical line denotes the end of the regular payment term. The vertical line in the middle denotes the point in time between $t^* - 1$ and t^* which we would use for defining default events in terms of profitability, i.e. where the profitability becomes negative. The decision rule that was finally implemented by the company providing the data is illustrated by the right vertical line. Here the end of the period of acceptable performance is set to $t^* - 1 = 27$, which is twelve months after the end of the regular payment term. The resulting classification rule for the default event Z is

$$z_i = \begin{cases} 1 & \text{if } \delta_{i,28} > 0 \\ 0 & \text{else} \end{cases} \tag{8}$$

The numbers of detected defaults and non-defaults on basis of the classification rule are given in Table 1.

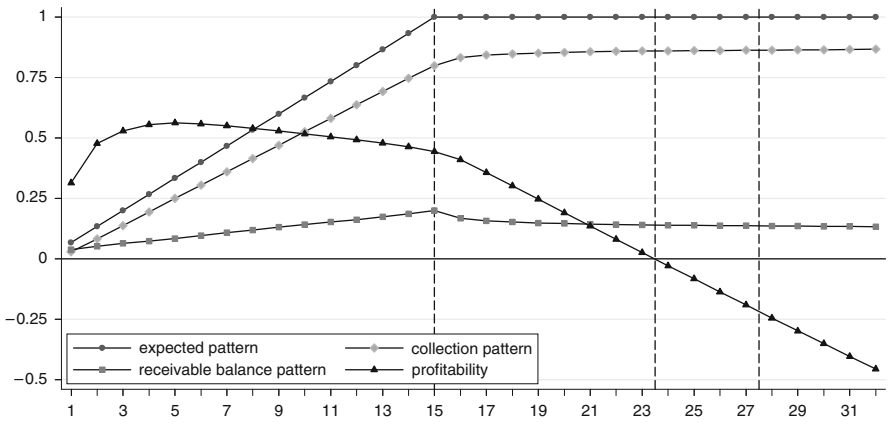


Fig. 1 Patterns of payment and profitability for the empirical data set

Table 1 Detected defaults and non-defaults for the empirical data set

Defaults ($z_i = 1$)	5 382	15.84%
Non-defaults ($z_i = 0$)	28 604	84.16%
Total	33 986	100.00%

4 Indicators of Individual Payment Performance

The proposed approach has two main disadvantages with respect to the monitoring and forecasting of credit default events in practice. First, the analysis of the payment patterns and the profitability of the portfolio of customers over time does not tell us anything about the individual performance of credits during the period of deficient payment ($t = 1, \ldots, t^*$). Second, this analysis is based on a retrospective analysis of complete histories which are observed for a very long period, even after the end of the agreed payment term. This is not suitable for monitoring and forecasting default events because such a data base is not available during the actual payment process where we observe payments in real time. Therefore, we propose two indicators for measuring the individual payment performance. The basic idea of these indicators is to evaluate the development of paid amounts with respect to the expected amounts and the time line of the payment process.

4.1 Description of Indicators

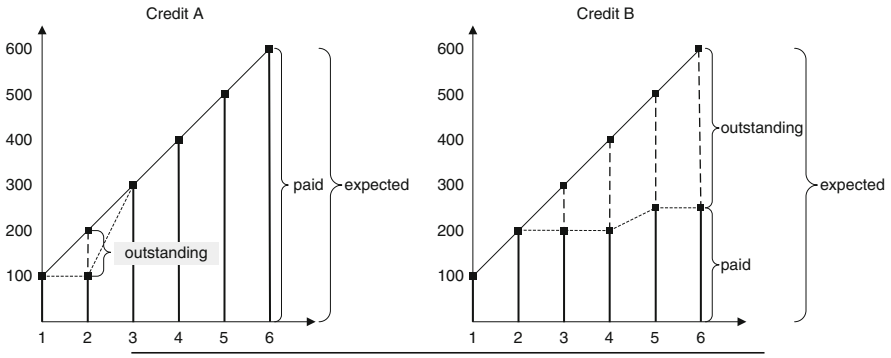
The first indicator $L_{i,k}$, which we call the individual liquidity, relates the cumulated amounts which are paid until k to the respective cumulated expected amounts:

$$L_{i,k} = \frac{X_{i,k}}{Y_{i,k}} \tag{9}$$

The individual liquidity at time k is therefore the proportion of due paid amounts until k. The second indicator $PC_{i,k}$ is called the individual payment career and relates the total stock of paid amounts to the total stock of expected payments:

$$PC_{i,k} = \frac{\sum_{t=1}^{k} X_{i,k}}{\sum_{t=1}^{k} Y_{i,k}} \tag{10}$$

Figure 2 gives two hypothetical examples of individual payment histories where a total financed amount of 600 Euros has to be paid off by six regular installments of 100 Euros each. Both payment histories show deficient payment. The customer who paid off credit A missed to pay at $t = 2$, but balances the account immediately at $t = 3$. The account related to credit B still shows deficient payment at the end of the regular payment term. From the respective figures of the performance indicators, the



	Credit A						Credit B					
$L_{i,k}$	1.00	0.50	1.00	1.00	1.00	1.00	1.00	1.00	0.67	0.50	0.50	0.42
$PC_{i,k}$	1.00	0.67	0.83	0.90	0.93	0.95	1.00	1.00	0.83	0.70	0.63	0.57

Fig. 2 Two examples of individual payment histories

main differences between the two indicators emerge. Whereas the individual liquidity equals 1 again at $t = 3$ for credit A, the individual payment performance keeps the deficient payment in mind for a longer time. But in contrast to the individual liquidity, the individual payment career can not directly be interpreted in financial terms, e.g. as proportion of due payments. Hence, it would be suitable to evaluate both indicators in parallel.

4.2 Application to the Empirical Data Set

For evaluating the defined indicators of individual payment performance on basis of the empirical data set we restrict their calculation to the regular payment term ($T = 15$). Figure 3 shows the box plots for the distribution of the individual liquidity (left) and the individual payment career (right) over $T = 15$. These plots exclude outside values. The data are grouped according to their definition as default and non-default given in Table 1. Obviously, both indicators show the ability to distinguish between good and bad accounts, and this ability is increasing with t. But the group-specific distributions of both indicators also show an area of intersection, even for large t.

To exemplify the early detection of default events we choose $L_{i,k} \leq 1/6$, that is, we calculate the number of defaults and non-defaults showing an individual liquidity of $L_{i,k} \leq 1/6$ for $k = 1, \ldots, T$. Using this cut-off value, we would detect 54.79% of the 5,382 credits defaults already at $t = 6$. For $t = 12$ this proportion increases to 57.62%, and to 61.93% at $t = 15$. At the same time, the proportion of non-defaults detected by this cut-off value for the individual liquidity decreases from 1.56% at $t = 6$ to 0.66% at $t = 12$ and to 0.50% at $t = 15$. Although especially the misclassification error is relatively low, it has to be remembered that this

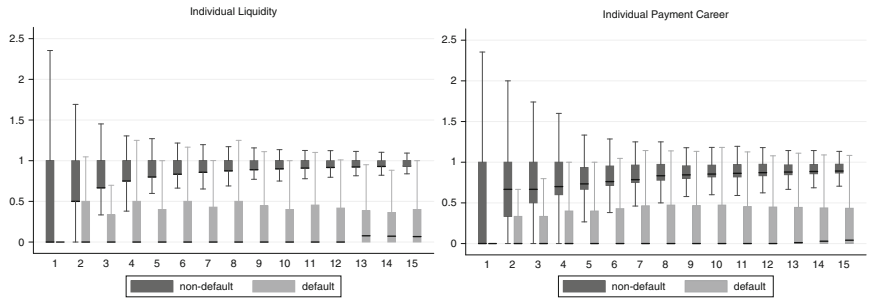


Fig. 3 Box plots of individual liquidity and individual payment career

classification directly depends on the definition of default in (8) and the underlying cost ratio when measuring the profitability. Furthermore, the analysis is based on empirical, historical data of credit histories, and this implies that the company took action to collect outstanding amounts. It has to be assumed that the analysis is biased as these actions do or do not have an effect on the individual payment behavior of customers.

5 Discussion

A consistent concept of credit default and the definition of credit default events does not exist. To contribute to the scientific discussion on these topics we proposed a payment-pattern approach to the definition of credit default events, and we defined and analyzed indicators of individual payment performance. Such indicators can be useful tools for monitoring payment behavior. The proposed measures and indicators should be improved by relating the payment histories to the respective histories of collection activities. This would directly bring up advice on how to improve accounts receivable management. Since this is not available so far, our further research in this area will concentrate on the comparative analysis of further appropriate methods, like transition probability and event history analysis.

References

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society, Series A, 160*, 523–541.

Johnson, R. W., & Kallberg, J. G. (1986). Management of accounts receivable and payable. In E. I. Altman (Ed.), *Handbook of corporate finance*, Section 8. New York: Wiley.

Stone, B. K. (1976). The payments-pattern approach to the forecasting and control of accounts receivable. *Financial Management, 5*, 65–82.

Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit scoring and its applications*. Philadelphia: SIAM.

Forecasting Candlesticks Time Series with Locally Weighted Learning Methods

Javier Arroyo

Abstract In finance, candlestick charts describe the price movements of an equity over time. Each candlestick summarizes the variability of a trading session by means of two pairs of values: the opening-closing and the lowest-highest prices of each session. In this sense, candlesticks can be seen as a kind of symbolic variable.

Candlestick charts are believed to reflect the psychology of the market and are used by technical analysts to make investment decisions. However, despite their popularity, little academic research has been done in order to determine if they are useful to forecast the future state of the considered equity. This article is devoted to this purpose as it proposes to forecast candlestick time series using locally weighted learning methods (Atkeson et al. 1997), such as the k-Nearest Neighbors algorithm. This kind of methods have been successfully applied to forecast other kind of financial time series (Aparicio et al. 2002; Fernández-Rodríguez et al. 1999). The forecasting ability of the proposed methods will be illustrated with a candlestick time series of the S&P500 stock index.

1 Introduction

A high frequency financial time series is a sequence of transaction prices of a given financial asset observed throughout time. This kind of time series may contain hundreds or thousands of prices for each single trading day. The analysis of these time series is not trivial as is complicated by irregular temporal spacing, diurnal patterns, price discreteness, and complex dependence (Engle and Russell 2009). Moreover, it is impossible to forecast the whole sequence of intradaily prices for the next day. As a result, financial forecasters usually deal with the time series of the daily close prices (or returns). However, in doing so, the valuable information contained in the rest of the intra-daily prices is neglected.

J. Arroyo
Dpto. de Ingeniería del Software e Inteligencia Artificial, Universidad Complutense de Madrid, Prof. José García Santesmases s/n 28040 Madrid (Spain)
e-mail: javier.arroyo@fdi.ucm.es

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_66,

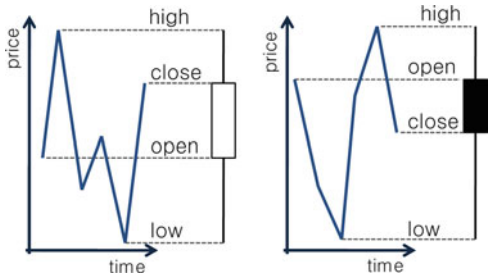


Fig. 1 White and black candlesticks

One approach to avoid this loss is to aggregate the intra-daily values by means of a symbolic variable. In other words, to summarize the high frequency financial time series by a lower frequency symbolic time series. Examples of this approach can be found in Arroyo et al. (2009), where interval time series represent the range between the daily lowest and highest prices of a financial asset, and in Arroyo and Maté (2009), where histograms are used to represent the distributions of intra-daily returns, giving rise to a daily histogram time series of returns. However, other kind of symbolic variables can be applied. In this work, candlesticks, a well-known financial tool, will be considered as a new kind of symbolic variable.

In financial newspapers and web sites, candlestick charts are used to represent the temporal evolution of asset prices. In these charts, each period of time is described by a candlestick composed of two meaningful intervals

- $[Low, High]$ that represents the volatility of the asset in the period
- $[Open, Close]$ that represents the return of the asset in the period

Figure 1 shows how candlesticks summarize price time series. The body is black if the opening price is greater than the closing price, and white otherwise. Candlestick charts offer an informative summary of the prices which is believed to reflect the psychology of market. Technical analysts study candlestick charts looking for patterns that help them to make profitable investment decisions. However, their heuristics often lack a rigorous basis. The fuzzy approach proposed in Lee et al. (2006) to represent candlesticks is a first step to alleviate this problem.

Despite the scarcity of rigorous approaches dealing with candlesticks, the four values that characterize them have been used for a long time in finance to estimate the volatility of an asset (Garman and Klass 1980; Rogers and Satchell 1991). This fact along with the study in Fiess and MacDonald (2002) prove that these four values provide valuable information.

Knowing these four values in advance can be very useful to make investment decisions and to determine the future volatility. Thus, the aim of this work is to propose a forecasting method for candlestick time series. A locally weighted learning (LWL) method will be adapted for this purpose.

LWL methods can be useful to estimate complex non-linear functions. As financial time series are supposed to be non-linear, it is expected to obtain good results with LWL methods. However, literature shows mixed evidence about this fact

(Aparicio et al. 2002; Fernández-Rodríguez et al. 1999; Meade 2002), which is not surprising due to the intrinsic difficulty of financial forecasting.

The next section is devoted to the adaptation of a LWL method to forecast candlestick time series.

2 Locally Weighted Learning Methods for Candlestick Time Series

Lazy learning is a form of learning that defers processing of training data until a new instance arrives. As it only involves storing the past instances, it is also known as memory-based learning. Lazy learning can deal with complex target functions and does not miss information in the training process. On the other hand, it can be easily fooled by irrelevant attributes and the response time for a new query is greater than for other learning approaches.

LWL is a form of lazy learning that combines in some way the most relevant instances from training data to yield a solution. In this kind of learning, instead of estimating a single global model for the whole data set, a new local model is estimated for each new instance using only the information provided by the closest already known instances. LWL methods include nearest neighbors, weighted averages and locally weighted regression. A comprehensive review of this field can be found in Atkeson et al. (1997).

In the next section, a LWL method, the k-Nearest Neighbors (k-NN), will be adapted to forecast candlestick time series. The adaptation of this method extends the adaptation for interval time series proposed in Arroyo (2008).

2.1 k-NN for Candlestick Time Series

Let $\{Q_t\}$ with $t = 1, \ldots, n$ be a candlestick time series (CTS) where each candlestick Q_t is defined by a quadruple of values: the open O_t, close C_t, low L_t, and high H_t prices at time t. Q_t can also be defined by the low-high interval $[LH]_t = [L_t, H_t]$, where $L_t \leq H_t$, and the open-close interval $[OC]_t = [O_t, C_t]$, without the $O_t \leq C_t$ constraint. The k-NN for CTS requires to represent $\{Q_t\}$ as a series of d-dimensional candlestick vectors $Q_t^d = (Q_t, Q_{t-1}, \ldots, Q_{t-d+1})$.

Given that, the k-NN consists in two steps: determination of the k nearest neighbors and generation of the forecast. They are shown next.

2.1.1 Determination of the k Nearest Neighbors

In order to determine the distance between the last candlestick vector Q_n^d and all the past candlestick vectors Q_t^d with $t = d, \ldots, n-1$, a distance for candlestick

vectors has to be defined. For this purpose, the kernel-based interval distance shown in González et al. (2004) will be extended to deal with candlesticks.

Before introducing this distance, it is needed to clarify that an interval $[X] = [X_L, X_U] = \langle X_C, X_R \rangle$ can be defined by its lower X_L and X_U upper bounds or, alternatively, by its center $X_C = (X_L + X_U)/2$ and its radius $X_R = (X_U - X_L)/2$. Given that, the kernel-based distance between intervals $[A]$ and $[B]$ is defined as

$$D_k([A],[B]) = \sqrt{(B_C - A_C)^2 + (B_R - A_R)^2}. \tag{1}$$

This distance is in fact the Euclidean distance between two intervals defined by their centers and their radii. More details about the properties of the distance and the kernel considered can be found in González et al. (2004).

If we consider that a candlestick is an individual defined by two intervals, then we can propose a Minkowski metric in $\Re^2$ using the distance in (1), similar to what is done with the Hausdorff distance for intervals in Palumbo and Irpino (2005). The resulting distance for two candlesticks $Q_i = \{[OC]_i, [LH]_i\}$ with $i = 1, 2$ and where $[LH]_i = \langle LH_{Ci}, LH_{Ri} \rangle$ and $[OC]_i = \langle OC_{Ci}, OC_{Ri} \rangle$ (being possible $OC_{Ri} < 0$) is given by

$$D_{Q,s}(Q_1, Q_2) = (D_k([OC]_1, [OC]_2)^s + D_k([LH]_1, [LH]_2)^s)^{1/s}, \tag{2}$$

where D_k is the distance in (1) and s is the order of the Minkowski metric.

Given the distance in (2), the dissimilarity between the last candlestick vector Q_n^d and all the past candlestick vectors Q_t^d with $t = d, \ldots, n-1$ can be represented as a distance-based Root Mean Square Error

$$RMSE_{Q^d,s}(Q_n^d, Q_t^d) = \sqrt{\frac{1}{d}\sum_{t=1}^{d} D_{Q,s}(Q_{n-i+1}, Q_{t-i+1})^2}. \tag{3}$$

Once all the $n - d$ distances are computed, the k closest vectors are identified and denoted as $Q_{t_p}^d$ with $p = 1, \ldots, k$.

2.1.2 Generation of the Forecast

The forecast $\hat{Q}_{n+1}$ has to be generated from the subsequent candlesticks of the k closest vectors, i.e., from Q_{t_p+1}. It is proposed that $\hat{Q}_{n+1} = \{\langle \hat{OC}_{Cn+1}, \hat{OC}_{Rn+1} \rangle, \langle \hat{LH}_{Cn+1}, \hat{LH}_{Rn+1} \rangle\}$ be a convex linear combination of the candlestick components with weights ω_p, where $\omega_p \geq 0$ and $\sum_{p=1}^{k} \omega_p = 1$,

$$\hat{OC}_{Cn+1} = \sum_{p=1}^{k} \omega_p OC_{Ct_p+1}, \quad \hat{OC}_{Rn+1} = \sum_{p=1}^{k} \omega_p OC_{Rt_p+1},$$
$$\hat{LH}_{Cn+1} = \sum_{p=1}^{k} \omega_p LH_{Ct_p+1}, \quad \hat{LH}_{Rn+1} = \sum_{p=1}^{k} \omega_p LH_{Rt_p+1}. \tag{4}$$

In the unweighted k-NN, $\omega_p = 1/k, \ \forall k$. While, in the weighted k-NN

$$\omega_p = \frac{\psi_p}{\sum_{l=1}^{k} \psi_l}, \text{ with } \psi_p = (D(Q_n^d, Q_{t_p}^d) + 10^{-8})^{-1}, \tag{5}$$

where 10^{-8} avoids the infinite values caused by zero distances.

3 Forecasting the S&P500 Candlestick Time Series

In this section, the proposed k-NN method will be applied to forecast the Standard & Poor's 500 (S&P500) CTS in 2007. The S&P500 is a stock index that includes the 500 companies with the largest capitalization in the United States. The required data can be downloaded from financial websites such as http://finance.yahoo.com/.

As it is well known, financial prices time series often show a stochastic trend. Unfortunately, the k-NN method does not work well on trended time series, because trends make it harder to find neighbors similar to the current sequence in the past of the time series. Thus, it is needed to propose approaches to remove the stochastic trend from the CTS.

3.1 Removing the Trend from Candlestick Time Series

In order to remove the stochastic trend from the time series of the four candlestick components, some kind of difference operator should be applied to render them stationary. Two approaches will be explored.

3.1.1 Differencing the Intervals

In Arroyo et al. (2009), a difference operator is proposed to remove the stochastic trend in interval time series. Given an ITS $\{[X]_t\} = \{[X_{Lt}, X_{Ut}]\} = \{\langle X_{Ct}, X_{Rt}\rangle\}$, the method consists in doing the following operation

$$\langle X_{Ct} - X_{Ct-1}, X_{Rt}\rangle = [X_{Lt} - X_{Ct-1}, X_{Ut} - X_{Ct-1}], \ \forall t. \tag{6}$$

This operation removes the stochastic trend from the interval positions, i.e., from the center time series.

This operation can be applied to the open-close and low-high intervals that compose the CTS, but, as both intervals are independently differenced, the candlestick property $[O_t, C_t] \subseteq [L_t, H_t]$ is not necessarily fulfilled for all t in the differenced CTS. This is not a problem because the original CTS is obtained once the difference operation is reversed. However, if the k-NN method is applied to forecast the differenced CTS, it cannot be guaranteed that the forecasted CTS fulfills this property once the difference operation is reversed. Thus, this approach will be ruled out.

3.1.2 Differencing the Candlestick Using the Previous Close Value

The stochastic trend can also be removed by subtracting to all the components of the present candlestick the previous close price C_{t-1}

$$[O_t - C_{t-1}, C_t - C_{t-1}, L_t - C_{t-1}, H_t - C_{t-1}], \ \forall t. \tag{7}$$

This approach fulfills the property $[O_t, C_t] \subseteq [L_t, H_t], \forall t$ and in most cases will remove the stochastic trend from the time series of the four candlestick components.

3.2 The One-Step Ahead Forecasting Experiment

The daily CTS in 2007 (see an extract in Fig. 2) has been divided into the initialization set (first 50 periods), the training set (next 150 periods) and the test set (last 51 periods). The initialization set is used to provide the k-NN with enough past to look for the neighbors. While the training set is used to determine the number of neighbors k and the length of the vector d applied in each k-NN. These values will be the ones that minimize the error in the training set. Error will be measured as the distance-based RMSE given by

$$RMSE_{Q,s}(\{Q_t\}, \{\hat{Q}_t\}) = \sqrt{\frac{1}{d}\sum_{t=1}^{n} D_{Q,s}(Q_t, \hat{Q}_t)^2}, \tag{8}$$

where $D_{Q,s}$ is a Euclidean-like distance obtained by taking the distance in (2) with $s = 2$.[1]

The candlestick k-NN method has been applied to two time series: the CTS and the close-value differenced CTS. For each of these time series, the k-NN has been applied with both the unweighted and the weighted schemes. The naïve

[1] The dissimilarity measure in (3) will also be estimated with $s = 2$.

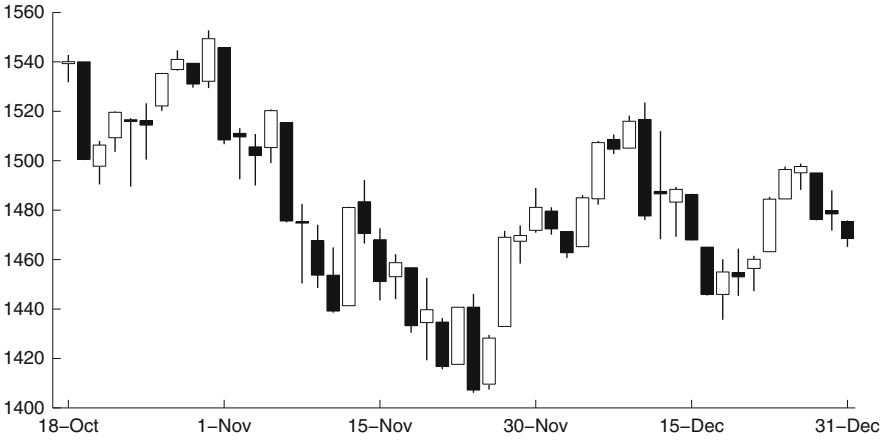


Fig. 2 Test period of the SP500 candlestick time series

Table 1 RMSE in each of the candlestick components time series in the test set

Method	Open	Close	Low	High
Naïve	17.7	19.8	14.6	14.5
Independent univar. k-NN unweighted	18.8	21	15.1	18.5
Independent univar. k-NN weighted	19.8	20.2	16.6	18.5
Independent diff. univar. k-NN unweighted	19.5	19.8	13.9	14.4
Independent diff. univar. k-NN weighted	17.4	**19.3**	13.9	15.2
Candle. k-NN unweighted ($k = 4$, $d = 1$)	4.7	21	13.9	12.8
Candle. k-NN weighted ($k = 5, d = 1$)	4.7	19.8	**13.4**	12.4
Close-value diff. candle. k-NN unweighted ($k = 20, d = 1$)	**3.1**	20.1	14.3	**11.4**
Close-value diff. candle. k-NN weighted ($k = 20, d = 1$)	**3.1**	20.1	14.2	11.5

method $\hat{Q}_{t+1} = Q_t$ has been used as the benchmark. In addition, to compare the candlestick k-NN with a non-symbolic approach, the univariate k-NN has been independently applied to forecast each one of the four candlestick component time series, both differenced and non differenced. The performance of all of these forecasting approaches is shown in Table 1 that displays the Root Mean Square Error in the test period for each one of the candlestick components time series.

The first conclusion that can be drawn from the table is that forecasting the CTS as a whole is much better than doing it as four independently-considered time series. In addition, the latter approach does not guarantee that the four independently forecasted values are a valid candlestick. These facts reinforce the idea that if data is symbolic, then they should be analyzed with symbolic methods.

The components where the improvement is more significant are the open and the high prices. On the other hand, in the close price no improvement with respect to

the naïve method is obtained. This is not surprising as it is very difficult to beat the naïve method when forecasting financial close prices.

Another interesting conclusion from the table is that the optimal length of the sequence for all the k-NN methods is $d = 1$, which means that only the candlestick in t is needed to characterize the time period t and that the past periods do not provide relevant information. This fact also concurs with financial theory, which considers that all the information relevant to the price can be found in the present period.

4 Future Work

The k-NN method proposed to forecast CTS improved the naïve method performance, which is the usual benchmark in financial time series, in all the components except in the close value, where is usually very difficult to improve the naïve method performance. However a question arises, are these results enough to obtain profits? A simulation with trading rules using the forecasted candlesticks should be done to clarify this point.

Other line of future work would be to improve the proposed method by explicitly taking into account the orientation of the open–close interval as is done for principal component analysis in Irpino (2006). The forecasting performance for other time horizons should also be analyzed. Moreover, other LWL methods, such as locally weighted regressions, can also be proposed to forecast CTS.

Acknowledgements I acknowledge support from the project *Agent-based Modelling and Simulation of Complex Social Systems* (SiCoSSys), supported by Spanish Council for Science and Innovation with grant TIN2008-06464-C03-01, and from the *Programa de Creación y Consolidación de Grupos de Investigación UCM-BSCH* (GR58/08). Finally, I thank to the IFCS for the Chikio Hayashi Award.

References

Aparicio, T., Pozo, E., & Saura, D. (2002). The nearest neighbour method as a test for detecting complex dynamics in financial series. An empirical application. *Applied Financial Economics, 12*(7), 517–525.

Arroyo, J. (2008). *Métodos de predicción para series temporales de intervalos e histogramas*. PhD thesis, Universidad Pontificia Comillas, Madrid.

Arroyo, J., Espínola, R., & Maté, C. (2009). Forecasting interval time series: A comparison in finance. *Computational Economics* (to appear).

Arroyo, J., & Maté, C. (2009). Forecasting histogram time series with k-nearest neighbours methods. *International Journal of Forecasting, 25*, 192–207.

Atkeson, C. G., Moore, A. W., & Schaal, S. (1997). Locally weighted learning. *Artificial Intelligence Review, 11*(1–5), 11–73.

Engle, R. F., & Russell, J. R. (2009). Analysis of high frequency financial data. In Y. Ait-Sahalia and L. Hansen (Ed.), *Handbook of Financial Econometrics, Tools and Techniques*, Amsterdam: North-Holland, *1*, 383–426.

Fernández-Rodríguez, F., Sosvilla-Rivero, S., & Andrada-Félix, J. (1999). Exchange-rate forecasts with simultaneous nearest-neighbour methods: Evidence from the EMS. *International Journal of Forecasting, 15*(4), 383–392.

Fiess, N. M., & MacDonald, R. (2002). Towards the fundamentals of technical analysis: Analysing the information content of high, low and close prices. *Economic Modelling, 19*(3), 353–374.

Garman, M. B., & Klass, M. J. (1980). On the estimation of security price volatilities from historical data. *The Journal of Business, 53*(1), 67–78.

González, L., Velasco, F., Angulo, C., Ortega, J. A., & Ruiz, F. (2004). Sobre núcleos, distancias y similitudes entre intervalos. *Inteligencia Artificial, Revista Iberoamericana de IA, 8*(23), 111–117.

Irpino, A. (2006). Spaghetti PCA analysis: An extension of principal components analysis to time dependent interval data. *Pattern Recognition Letters, 27*(5), 504–513.

Lee, C.-H. L., Liu, A., & Chen, W.-S. (2006). Pattern discovery of fuzzy time series for financial prediction. *IEEE Transactions on Knowledge and Data Engineering, 18*, 613–625.

Meade, N. (2002). A comparison of the accuracy of short term foreign exchange forecasting methods. *International Journal of Forecasting, 18*(1), 67–83.

Palumbo, F., & Irpino, A. (2005). Multidimensional interval-data: Metrics and factorial analysis. In *Conference on Applied Statistical Models and Data Analysis (ASMDA)* (pp. 689–698).

Rogers, L. C. G., & Satchell, S. E. (1991). Estimating variance from high, low, and closing prices. *The Annals of Applied Probability, 1*(4), 504–512.

An Analysis of Alternative Methods for Measuring Long-Run Performance: An Application to Share Repurchase Announcements

Wolfgang Bessler, Julian Holler, and Martin Seim

Abstract Measuring the long-run financial performance around and subsequent to specific corporate announcements is important and of special interest to researchers and practitioners alike. A variety of methods have been developed and applied to estimate these valuation effects but so far there is no general consensus on the best approach from a financial and statistical perspective. The objective of our research is to analyze and compare empirically alternative methods using 144 repurchase announcements in Germany for the period from 2000 to 2006. Overall, we find that the methodology used strongly influences the results.

1 Introduction

One of the major research areas in corporate finance are the long-run valuation effects subsequent to specific corporate events and especially financing decisions such as dividend changes and share repurchases. Although there exists a large number of empirical studies there is hardly any consensus on the best approach for measuring long-run performance. In addition to the specification of risk factors and the appropriate benchmark there are a number of methodological issues. For instance, many tests are often biased due to event clustering and cross-sectional correlations in abnormal performance measures. While new approaches have been developed to address these problems, most suffer from new biases and only allow for limited cross-sectional inference. In order to address these problems, we implement a new approach developed by Höchle et al. (2009) that attempts to overcome these shortcomings and compare the results to those of alternative approaches. We apply various methods to a dataset of 144 share repurchase announcements in Germany for the period from 2000 to 2006. During the last decade, share repurchases have become an important alternative to dividends for distributing cash flows

M. Seim (✉)
Center for Finance and Banking, Justus-Liebig University, Licher Strasse 74, 35394 Giessen
e-mail: martin.seim@wirtschaft.uni-giessen.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_67,

to shareholders. As documented in the literature, share repurchases often result in substantial short- and long-run valuation effects which can be explained by two competing hypotheses (Bessler et al. 2009). The 'signaling'-hypothesis proposes that stock repurchases convey positive private information of corporate managers to the stock markets. In contrast, the 'free cash flow'-hypothesis views the distribution of cash flows to shareholders as a means to control managerial moral hazard (Bessler et al. 2010). Overall, we find significant stock price reactions for all different methods. However, the magnitude and the direction of the results depend on the method employed.

2 Methodologies for Measuring Long-Run Performance

Various methods such as buy-and-hold abnormal returns (BHAR) and the calendar time approach have been developed for measuring the valuation effects and the long-run financial performance subsequent to corporate announcements and events. In order to overcome the statistical shortcomings of these methods, Höchle et al. (2009) developed a generalized calendar time approach. All these methods differ with respect to their statistical assumptions and the cross-sectional factors used for explaining abnormal stock returns. The intuition as well as the strengths and weaknesses of the three approaches are analyzed in the following sections.

2.1 BHAR, Fama-French Alphas and Cross-Sectional Regressions

The approach most often used for analyzing long-run abnormal stock returns is to determine buy-and-bold abnormal returns (BHAR). These are calculated for company i as follows:

$$BHAR_i = \left(\prod_{t=1}^{T}(1 + R_{i,t})\right) - \left(\prod_{t=1}^{T}(1 + R_{M,t})\right) \tag{1}$$

where T indicates the holding period, $R_{i,t}$ is the stock's return and $R_{M,t}$ is the return of a market index used as a benchmark. This performance measure compares the average performance of a buy-and-hold strategy of investing in all companies at the event date to a buy-and-hold investment in a broad-based market benchmark with a similar risk profile as the event companies. In order to control for the impact of common risk factors, some studies employ a time-series regression for each event company on the Fama-French-factors *SMB* and *HML*:

$$R_{i,t} - R_{f,t} = \alpha_i + \beta_M \left(R_{M,t} - R_{f,t}\right) + \beta_{SMB} R_{SMB,t} + \beta_{HML} R_{HML,t} + \epsilon_i \tag{2}$$

where α_i measures the outperformance and *HML* (high minus low) and *SMB* (small minus big) represent the factor mimicking portfolios. In order to determine the relevant economic variables that explain the abnormal stock performance, the cross-section of the BHAR and alpha estimates are regressed on a set of explanatory variables:

$$BHAR_i = \alpha + \beta \cdot x_i + \epsilon_i \quad \text{or} \quad \alpha_i = \alpha + \beta \cdot x_i + \epsilon_i \tag{3}$$

where x_i is a vector containing company-specific variables and ϵ_i is an $i.i.d.$ error-term. Despite its intuitive appeal and its ability to include many different types of explanatory variables in the cross-section, this approach suffers from one major drawback. In many corporate finance events there is a pronounced clustering of events over time. With the simple one-factor model used in most BHAR applications, this may result in cross-sectional dependence in abnormal performance estimates due to omitted risk factors which often leads to underestimated standard errors. This problem is most pronounced for longer holding periods and is more severe for a higher frequency of overlapping events (of the same firm or across firms).

2.2 Calendar Time Portfolio Approach

The problems of event-clustering and cross-sectional correlations are addressed the calendar time portfolio approach. This method takes advantage of the fact that portfolio formation naturally adjusts for any cross-sectional dependence in stock returns. Thus, equally weighted portfolios are formed in calendar time. Companies are included for a predetermined holding period which begins at the event date. More formally, for each time period t an equally weighted average of all stock returns is calculated:

$$R_{P,t,l} = \frac{1}{N_t} \sum_{i=1}^{N} p_{i,t,l} \cdot R_{i,t} \tag{4}$$

where N is the total number of firms in the sample and $p_{i,t,l}$ is an indicator variable equal to one if the firm had an event during the last months where l defines the holding period. N_t equals the number of firms for which $p_{i,t,l}$ equals 1. To determine the abnormal performance, a time-series regression on a given set of risk factors such as the Fama-French factors is performed:

$$R_{P,t} - R_{f,t} = \alpha_P + \beta_M \left(R_{M,t} - R_{f,t}\right) + \beta_{SMB} R_{SMB,t} + \beta_{HML} R_{HML,t} + \epsilon_i \tag{5}$$

A positive and significant α indicates an outperformance. However, this approach has a weakness in that it imposes limits on the cross-sectional analysis of the variables that explain abnormal performance. In fact, it is restricted to the analysis of dichotomous variables. From a statistical point of view, this approach is biased in case of event-clustering, because it puts excessive weight on those events that take place during time periods with a low frequency of events. There are additional

problems in the application of this approach. First of all, the length of the holding period has an impact as different events might influence stock returns for different time periods. It further determines the stability of the composition of the calendar time portfolios. This becomes an important issue if there is event-clustering. In addition, some companies may have multiple events so that holding periods for the same company might overlap. In this case the calendar time portfolio puts a relatively high weight on individual companies for specific intervals.

2.3 Generalized Calendar Time Approach

In order to combine the statistical advantages of the calendar time approach with the cross-sectional analysis, Höchle et al. (2009) developed a 'Generalized Calendar Time Approach' (GCT approach). They show that a pooled panel regression augmented by Driscoll and Kraay (1998) standard errors is able to replicate the results of the calendar time approach for the case of a dichotomous variable. The panel structure of this model naturally carries over to other discrete as well as continuous variables. The structure of their panel regression model is given by:

$$R_{i,t} = [(p_{i,t} \otimes x_{i,t}) \otimes z_t] \cdot \beta + \upsilon_{i,t} \tag{6}$$

where z_t denotes the set of risk factors specified by the specific asset pricing model. These can vary over time and are the same for all firms; $x_{i,t}$ is a set of firm characteristics which can vary over time as well as across firms; $p_{i,t}$ contains a constant and a dummy variable indicating whether an event took place over a specific time period. The significant innovation is the application of nonparametric Driscoll-Kraay standard errors which account for spatial correlations and therefore corrects for the major shortcoming of the BHAR approach while still allowing for cross-sectional inference.

3 Data and Empirical Results

For the empirical analysis, we use a sample of 144 share repurchase announcements that occurred between December 2000 and September 2006. 40 announcements were made by established firms listed in the DAX or MDAX and 104 by start-up firms that went public at the 'Neuer Markt'. For all methods we limit our analysis to this specific time period to ensure that all calculations are based on the same events. A holding period of 24 months is used, beginning in the month subsequent to the repurchase announcement. In order to compare the capability of the different approaches for determining the relevant variables that may explain the abnormal performance, we selected the following firm-specific control variables: debt-to-asset ratio (LEV), return on equity (ROE), market value (MV), market-to-book ratio

(MTB), and cash-to-assets ratio (CTA). The factor mimicking Fama-French portfolios were calculated by using the MSCI Germany style indices for small and large caps as well as the corresponding indices for value and growth stocks. In addition, the CDAX performance index is used as a market index and the EURIBOR 3-month rate represents the risk-free rate.

3.1 BHAR, Fama-French Alphas and Cross-Sectional Regressions

Table 1 indicates an outperformance of nearly 20% for DAX and MDAX firms over the 2 year period following the repurchase announcement based on the BHAR approach. For this subset the BHAR are significantly different from zero for each time horizon. In contrast, the BHAR for IPOs are insignificant, although there appears to be a positive trend. However, Fama-French alpha estimates are on average significantly different from zero for both the DAX and MDAX firms as well as for the IPO subsample. Nevertheless, they are higher in magnitude for the IPO group.

The goal of the cross-sectional analysis is to explain BHAR and alpha estimates using accounting measures from the financial statement prior to the repurchase announcement. As reported in Table 2, these control variables fail to explain BHAR and alpha estimates in nearly all cases. However, when alpha or BHAR are used as the dependent variable in the regression, the coefficient estimates of the Fama-French alpha regressions (Panel B) have a lower magnitude compared to the BHAR regressions (Panel A). It needs to be recognized, however, that the alpha measures an average monthly return whereas the BHAR are estimated over the entire 24 months holding period. From an economic perspective, it is important to note that CTA has a positive and significant impact on the alpha estimate for IPO firms. This result is in line with the 'free cash flow'-hypothesis. When young firms are subject to severe agency problems due to an excessive cash position, the decision to payout additional cash flows should reduce managerial moral hazard. Hence, a repurchase announcement results in superior future firm performance. Moreover, it is fair to conclude that the variables explain at least part of the performance as the constant terms in the regressions are insignificant indicating diminishing performance.

Table 1 Single stock performance measures: 24 months BHAR and Fama-French alphas

	av. 24 months BHAR	av. 24 months Fama-French α
DAX/MDAX	**19.70%*****	**12.72%***
IPO NM	10.91%	**18.48%*****

* 10% significance level
*** 1% significance level

Table 2 Cross-sectional regressions: 24 months BHAR and Fama-French alphas

	Const.	LEV	ROE	MV	MTB	CTA
			Panel A: BHAR24			
DAX/MDAX	0.2337	0.0005	0.0141	−0.0000	−0.0811	0.0005
IPO NM	0.0901	0.0001	−0.0008	0.0002	−0.0618	0.0028
			Panel B: alpha24			
DAX/MDAX	0.0024	−0.0000	0.0000	−0.0000	0.0030	0.0001
IPO NM	−0.0038	0.0001	−0.0000	−0.0000	−0.0003	**0.0002***

* 10% significance level

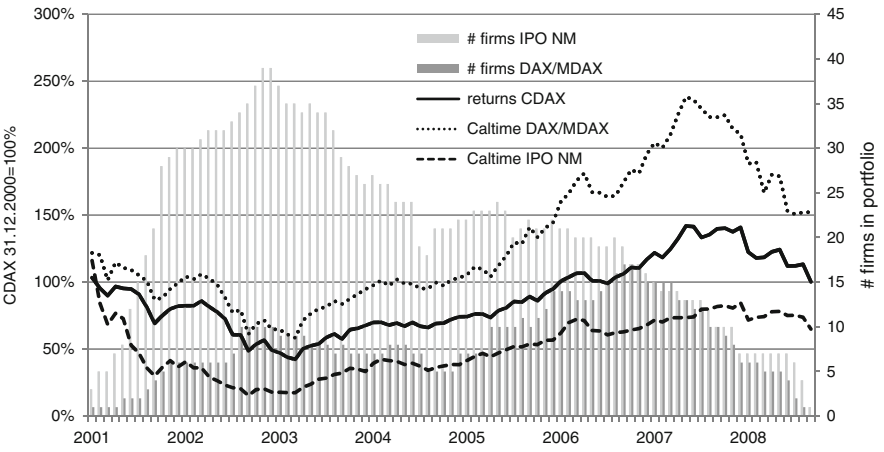


Fig. 1 Calendar time returns, benchmark and portfolio size for the 2001–2008 period

3.2 Calendar Time Portfolio Approach

The calendar time portfolios are constructed for a holding period of 24 months giving equal weight to each firm at each point in time. The number of firms in the portfolio and the portfolio returns for DAX/MDAX and IPO firms and the CDAX index are presented in Fig. 1.

It becomes immediately evident that our calendar time portfolios for both subsamples are biased towards events occurring during time periods when the event frequency is relatively low, i.e. at the beginning and at the end of the sample period. Table 3 reports the results for the Fama-French regressions of established firms and IPOs. DAX and MDAX firms exhibit a significant market exposure close to one while they have no significant loadings on *SMB* and *HML*. Finally, the positive intercept term indicates outperformance of the DAX/MDAX portfolio, albeit indistinguishable from zero. In contrast, the IPO portfolio is more strongly exposed to the market factor and significantly biased towards small stocks and growth stocks. The coefficient estimate for the intercept term is negative but statistically insignificant.

Table 3 Fama-French performance evaluation of calendar time portfolios

	α	Market	SMB	HML
DAX/MDAX	0.0035	**0.9933*** **	0.2875	−0.0939
IPO NM	−0.0010	**1.2156*** **	**0.8132*** **	**−0.6416*** **

*** 1% significance level

Table 4 GCT model: interaction of event firms with firm-specific and risk factors

	Panel A: Interaction Terms with Firm-Specific Variables					
	Event dummy	LEV	ROE	MV	MTB	CTA
DAX/MDAX	−0.0074	−0.0000	0.0003	−0.0000	0.0019	0.0001
IPO NM	**−0.0261** **	0.0000	**−0.0002*** **	−0.0000	**0.0077*** **	0.0003

	Panel B: Interaction Terms with Market Wide Factors		
	Market	SMB	HML
DAX/MDAX	0.0318	0.442	**0.4391*** **
IPO NM	**0.3518** **	0.2212	−0.3304

** 5% significance level
*** 1% significance level

Overall, the results for the DAX/MDAX events are in line with BHAR results and Fama-French alphas. In contrast, the negative alpha estimate of the IPO portfolio stands in contrast to the positive average alpha on a single event basis and also to the positive but insignificant 24 months BHAR.

3.3 Generalized Calendar Time Approach

Using the whole CDAX universe with stock and balance sheet data for more than 700 firms as a control sample, we estimate the GCT model in the last step. Table 4 presents the coefficient estimates only for the event dummy which equals one for the 24 months subsequent to the repurchase announcement as well as the interaction terms with the firm-specific variables and the risk factors.

Surprisingly, the coefficient estimate for the DAX/MDAX event dummy is negative but insignificant and hence weakly indicates an inferior performance of established firms that announced a share repurchase compared to the control group. While DAX/MDAX event firms do not exhibit any differences compared to the control group regarding the firm-specific variables, this group has a significantly stronger exposure to the value factor. In contrast, the event dummy for IPO firms is negative and significant corroborating the result from the Fama-French calendar time regression. Furthermore, a higher market-to-book ratio and a lower return on equity lead to an increase in the performance of IPOs, but these firms also carry a higher market exposure indicating that they are perceived as high risk firms.

4 Conclusion

Overall, we find empirical evidence that the results of long-run performance studies depend strongly on the method used. The GCT model indicates a significant underperformance of IPOs and a negative but insignificant performance for DAX/MDAX firms after controlling for firm-specific variables as well as market wide risk factors. However, the results are not in the same direction as the results of the cross-sectional regressions on BHAR and Fama-French alphas and the calendar time approach. These differences in empirical results are presumably closely related to clustering of repurchase activities in specific time periods. More precisely, event clustering induces cross-sectional correlations in the BHAR analysis but it also causes substantial variation in portfolio size as well as size differences between IPO and DAX/MDAX portfolios in the calendar time portfolio analysis. Thus, the methodology used strongly influences the results.

References

Bessler, W., Drobetz, W., & Seim, M. (2009). Motives and valuation effects of share repurchase announcements in Germany: A comparison of established firms and initial public offerings. Working Paper, University of Giessen.

Bessler, W., Drobetz, W., & Seim, M. (2010). Financing activities and payout policies of entrepreneurial firms: Empirical evidence from German initial public offerings. Working Paper, University of Giessen.

Driscoll, J., & Kraay, A. (1998). Consistent covariance matrix estimation with spatially dependent panel data. *Review of Economics and Statistics, 80*, 549–560.

Höchle, D., Schmid, M., & Zimmermann, H. (2009). A generalization of the calendar time portfolio approach and the performance of private investors. Working Paper, University of Basel.

Knowledge Discovery in Stock Market Data

Alfred Ultsch and Hermann Locarek-Junge

Abstract This work presents the results of a Data Mining and Knowledge Discovery approach on data from the stock markets using Databionic techniques. Stock market data is analyzed using methods that were learned from nature and previously applied primarily to DNA microarray data. It is demonstrated that the discovery of new insights into the stock markets is possible by the application of sensible preprocessing of daily returns (Relative Differences), application of a projection which has the potential to show emergent structures in the data (U-Matrix) and allows for a nontrivial clustering of the data (U*C).

1 Introduction

An issue that is the subject of intense debate among academics and financial professionals is the Efficient Market Hypothesis (EMH). It states that security prices fully reflect all available information at any time. The implications of the EMH are truly profound. Most individuals that buy and sell stocks in practice however, do so under the assumption that the securities they are buying are worth more than the price that they are paying, while securities that they are selling are worth less than the selling price.

Empirical evidence has been mixed, but has generally not supported strong forms of the efficient markets hypothesis, e.g. low P/E stocks have greater returns. Earlier papers also refuted the assertion that higher returns could be attributed to higher beta, which has been accepted by efficient market theorists as explaining the anomaly in neat accordance with modern portfolio theory. One can also identify "losers" as stocks that have had poor returns over some number of past years. "Winners" would be those stocks that had high returns over a similar period. Some trading rules say that in trends one should buy "winners" and sell "losers". While proponents of the EMH don't believe that it is possible to beat the market, some

A. Ultsch (✉)
Databionics Research Group, University of Marburg, Germany
e-mail: ultsch@informatik.uni-marburg.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_68,

believe that stocks can be divided into categories based on risk factors. However, these risk factors are considered to be stable over time. In this paper, we analyze a very large stock market to find out whether there exist groups of stocks and clusters of time, where the groups that we find behave similar in the way that the probability of rising or falling stock prices within the created groups can be forecasted and is different from randomness, which would challenge the EMH.

2 Daily Returns on Stocks

Primary data in this paper are the adjusted daily closing prices of stocks traded in the USA. The prices of 7031 stocks were collected from Yahoo Finance (finance.yahoo.com) for the period Jan. 1^{st} 2000 to 1^{st} of march 2008 (observation period). This resulted in 2047 trading days. A total of 14,390,410 stock prices were obtained in this way. Standard & Poor's 500 Index – S&P 500 gives an overall picture of the market situation during the observation period (see Fig. 1). The S&P 500 is one of the most commonly used benchmarks for the overall U.S. stock market. It can bee seen that the observation period rising as well as falling market conditions.

For each day (t) and each price p(t) the daily return was calculated as Relative Difference (*RelDiff(t)*):

$$RelDiff(t) = 2 * (p(t) - p(t-1))/(p(t) + p(t-1)$$

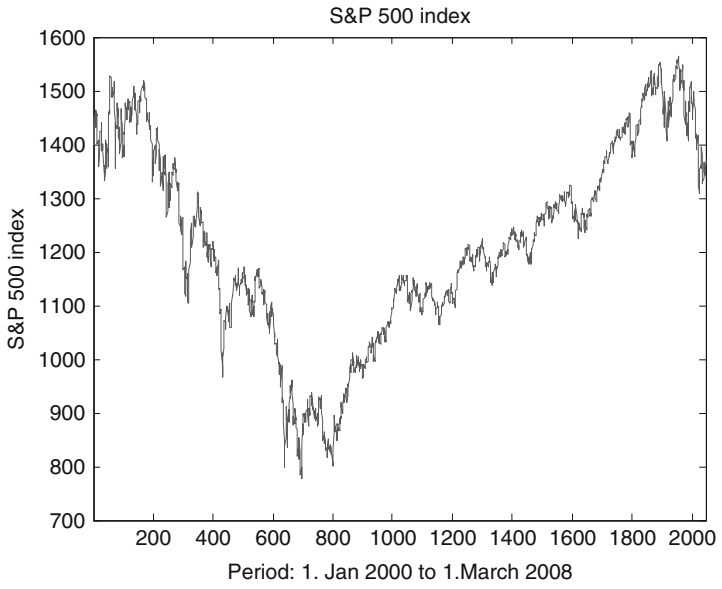


Fig. 1 S&P 500 during observation period

Relative Difference has several advantages over other formulas for return, like LogRatio ($\log(p(t)/p(t-1))$ or Ratio $(p(t)-p(t-1))/p(t-1)$. See Ultsch (2009) for a detailed discussion. The most important for this investigation is that RelDiff possesses a symmetric and finite range: if a company defaults ($p(t)=0$) then $RelDiff = -200\%$. If a company has exorbitant gains ($p(t-1) < p(t)$) then RelDiff approaches $+200\%$. This allows to model returns with finite variances. In Ultsch (2009) it was shown that returns measured in RelDiff can be modeled with a mixture of distributions using one Normal (Gaussian) and two LogNormal distributions. The definition of logarithms was generalized to negative numbers as $\log'(x) = sign(x) \cdot \log(abs(x))$. An initial LogNormal, Gaussian, LogNormal (LGL) model was fitted to the data using the Expectation Maximization algorithm (e.g. Izenman 2008). Figure 2 shows the empirical probability distribution measured with a kernel density estimator Pareto Density Estimation (PDE) (Ultsch 2003). The LGL model is depicted in Fig. 2 using dashed lines for each component and a solid line for the mixture. The quality of the model was assessed with a quantile/quantile plot resulting in an extremely good fit (see Ultsch 2009, Fig. 5).

This model can be naturally interpreted as a random result for returns, i.e. the central Gaussian $N(0, 1.7)$ with a fraction of 75% of all returns. Furthermore there are two non random distributions for returns, losses (12.5%) and wins (12.5%), which are lognormal distributed.

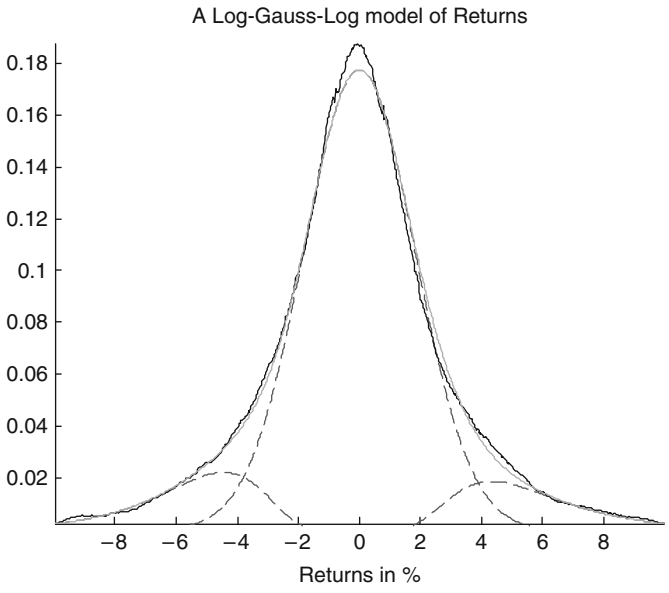


Fig. 2 The Log-Gauss-Log model of all returns

3 Knowledge Discovery in Market Activities

Using the model described in the last chapter it can be decided, whether a return belongs to the Random, Losses or Wins class using Bayes decision. We define $UnitWin = p(Return > Random) - p(Return < Random)$, where the probabilities are calculated with Bayes' theorem on the model developed above. UnitWin gives -1 for Losses, 0 for Random and $+1$ for Wins. In Fig. 3 UnitWin is shown for all returns.

The advantage of UnitWin is that differences in returns within same group are zero. UnitWin is therefore a good measure to compare the performance of different stocks for all trading days. The market activity on each day can be measured as the average number of non random returns for that day. This gives

$$Activity(t) = \underset{i}{mean}\,(abs(UnitWin(t,i)))$$

The distribution of Activity is shown in Fig. 4 using PDE.

It can be seen that Activity can be modeled as a mixture of Gaussians GMM (see Fig. 4). Using this GMM active days and inactive days can be distinguished. We found that the market was active for 2,045 days during our observation period. The next question is, whether there are days with more than average performance of the stock market. We defined the DailyPerformance(i) of a as the sum of all UnitWins for stock i. We found that the DailyPerformance consisted of three different distributions: a Gaussian around zero, i.e. passive performance or sideways

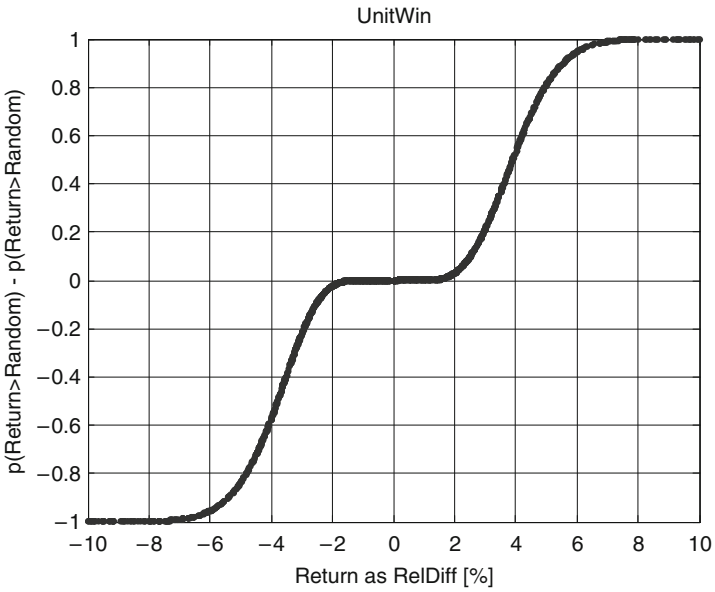


Fig. 3 UnitWin as a function of stock's returns

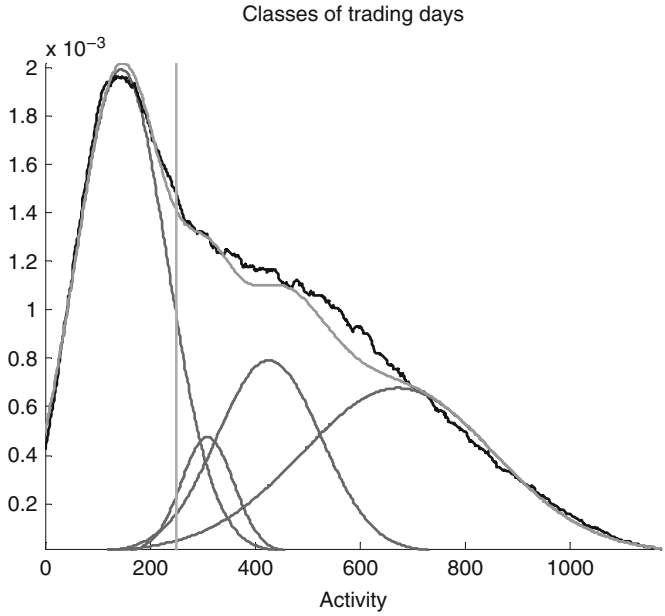


Fig. 4 Distribution of market activity for all days

movement of stocks and a winner and loser distribution, both lognormal distributed. Furthermore we found that only $568 = 8\%$ of all stocks dominated the performance of the stock market (marked leaders).

4 Types of Marked States

With UnitWins returns can be compared for sets of stocks and groups of days. The similarity respectively dissimilarity of marked days was defined as the Euclidean distance of the UnitWins of a set of stocks. Using this distance definition the different types of market days (winning, losing and passive) were compared for the marked leaders. For each of these groups a clustering procedure was performed using Emergent Self Organizing Feature Maps (ESOM) with the U-Matrix display (Deboeck and Ultsch 2000) and the clustering algorithm U*C (Ultsch 2007). Figure 5 shows an example of a U-matrix.

This 3D landscape is interpreted as follows: data in valleys are close in the high dimensional input space. Data separated by mountains are in different clusters. The $U * C$ clustering resulted in three clusters for Winner days $(w_1, \ldots w_3)$, four classes for Loser days$(l_1, \ldots l_4)$ and only one class for the Passive days.

As a next step the transition frequencies for each class were counted. The results is shown in Fig. 6. It is remarkable, that some states are rather persistent. For class l3, one of the loser classes, the probability that the next day is also a loser class is 74%.

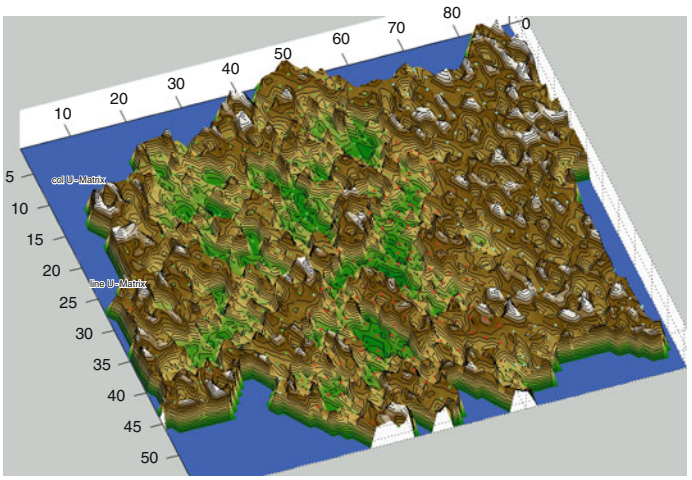


Fig. 5 U-Matrix of the winner days for the market leaders

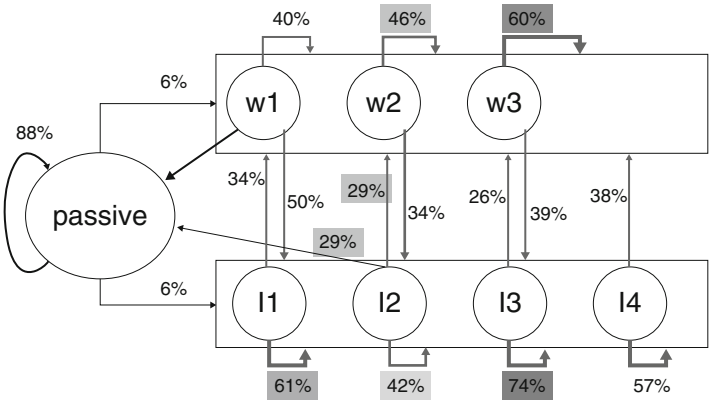


Fig. 6 Transition frequencies for the market classes

For class w_3 it was also observed that with a probability of 60% the next day is also a winning day. Other states are instable. E.g. in loser state l_2 with 58% probability, the next day is either passive or winning.

5 Discussion

This paper is an example of knowledge discovery in stock marked data. Knowledge Discovery is defined as the discovery of understandable knowledge which is new and useful. We have found that there are three types of returns: random, losses

and wins. Using Bayesian decision a meaningful aggregation of collective behavior could be defined (UnitWin).

Market activity was found to be either active or inactive. Daily performance could be classified in passive, winning and losing. UnitWin can be used for the definition of a sensible distance function. It has the advantage that inner group differences, e.g. within passive stocks or days, are zero.

Inter group differences contribute a precise and finite value to the distance function. Using this distance function, a clustering of the winner and loser market days was possible. The usefulness of these clusters can be seen in the transition frequencies to other. Some of the states suggest the buying (e.g. l_2) others the selling of stocks (e.g. w_1). It was not intended that this works may be used for the generation of buy-or-sell signals. It may, however, be useful to calculate measures for the overall state of a market day.

6 Conclusion

The EMH is the backbone of classical capital market theory. It has been tested empirically quite often, using econometrical testing and event studies. Several anomalies have been found, but they could mostly explained by applying risk measures and models for investor utility.

In this paper, knowledge discovery in stock marked data is applied. In the paper we found that there are three types of returns: random, losses and wins. A meaningful aggregation of collective behavior was defined and market activity was found to be either active or inactive while performance could be classified in passive, winning and losing. A clustering of the winner and loser market days was possible, where some of the states suggest buying, others the selling of stocks. It was not intended that this work may be used for the generation of buy-signals or sell-signals. It may, however, be useful to calculate measures for the overall state of a market day.

The authors will try to test the properties out-of-sample and in various other markets to find out whether the method works only in the sample period or it is a general property of the stock market, which remains to be proven with an independent test set. This challenge for the EMH remains future work.

References

Deboeck, G. J., & Ultsch, A. (2000). *Picking stocks with emergent self-organizing value maps* (Vol. 10, pp. 203–216). Prague: Neural Networks World, Institute of Computer Science.

Izenman, A. J. (2008). *Modern multivariate statistical techniques: regression, classification, and manifold learning*. Berlin: Springer.

Ultsch, A. (2003). *Pareto density estimation: A density estimation for knowledge discovery*. D. Baier and K. D. Wernecke (Eds.), Innovations in classification, data science, and information

systems – *Proceedings 27th annual conference of the german classification society (GfKl)* (pp. 91–100). Berlin, Heidelberg: Springer.

Ultsch, A. (2007). Analysis and practical results of U*C clustering. *Proceedings 30th annual conference of the german classification society (GfKl 2006)*. Berlin, Germany.

Ultsch, A. (2009). Is log ratio a good value for measuring return in stock investments? In: A. Fink, B. Lausen, W. Seidel & A. Ultsch (Eds.): *Advances in Data Analysis, Data Handling and Business Intelligence*, (pp. 505–511). Springer.

The Asia Financial Crises and Exchange Rates: Had there been Volatility Shifts for Asian Currencies?

Takashi Oga and Wolfgang Polasek

Abstract We analyse the volatility structure of Asian currencies against the U.S. dollar (USD) for the *Thai Baht* THB, the *Philippine Peso* PHP, the *Indonesian Rupiah* IDR and the *South Korean Won* KRW. Our goal is to check if the characteristics of the volatility dynamics have changed in a K-state AR(1)-GARCH(1,1) model in the last decade 1995–2008 covering the Asian crisis. We estimate the model of Haas et al. (2003) with MCMC and we find that for the four currencies the volatility dynamics has changed at least once.

1 Introduction

GARCH (generalized autoregressive conditional heteroscedasticity) models of Bollerslev (1986) have become very popular in econometrics to analyze the volatility structures of financial time series. Since the pioneering work of Hamilton (1989), Markov switching models have become a primary tool to analyze break points in time series. The last decade have seen some financial crises and it is interesting to see if these crises can be detected or are reflected in the volatility structure of exchange rates.

The term financial crisis is applied broadly to a variety of situations in which some financial institutions or assets suddenly loose a large part of their value. The consequences of financial crises can be manifold like banking panics, recessions and currency revaluations or system changes. Other situations that are often called financial crises include stock market crashes and the bursting of financial bubbles, as well as international phenomena like currency crises and sovereign defaults.

In the following we concentrate on volatility changes of four Asian currencies in the period 1995–2008, which covers the Asia financial crisis of 1997. Recall that the Asian financial crisis

T. Oga (✉)
Chiba University, 1-33 Yayoi-Cho, Inage-Ku, Chiba, 263-8522, Japan
e-mail: ohga@le.chiba-u.ac.jp

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_69,

Table 1 Asian currencies that changed from pegged to floating

Thailand	July 2, 1997
Philippine	July 11, 1997
Indonesia	August 14, 1997
South Korea	December 16, 1997

- Has started May 1997 in Thailand,
- Had most effects for Thailand, Indonesia and South Korea,
- Minor effects for Hong Kong, Malaysia, Laos and Philippines,
- While China, India, Taiwan, Singapore, Vietnam and Japan were less suffering.

The reason why we concentrate on these four currencies is the fact that these currencies gave up their currency pegs in the aftermath of the Asia crisis. Table 1 lists the dates when the four countries changed to a floating system. Note that these four changes occurred in the second half of the financial crisis year 1997. Singapore, China, Hong Kong and Russia did not change their currency systems following the Asia crisis. Similarly, their currencies were pegged mainly to the USD. Now our question is: Had the changes in the currency systems been accompanied by similar patterns in the volatility of the returns? If the pegs were abandoned because of speculative attacks, then the time point of the peg change must coincide with a break point in a volatility model of the currency returns. Can econometric models detect the regime shifts in the volatilities and do the estimated results correspond to the official dates? Analysing regime shifts in the volatilities since 1995, an econometric models could possibly detect other change points that were not necessarily related to the Asian crisis and might have been there for other reasons. We consider the data for the four currencies that are listed in Table 1: Thailand Baht THB, Philippine Peso PHP, Indonesia Rupiah IDR, and South Korea Won KRW from Jan. 3rd 1995 to mid 2008. We construct a Markov switching AR(1)-GARCH(1,1) model to analyze the structural change in the volatility dynamics and to interpret why the volatility change occurred. From a Bayesian view, we construct a *Markov chain Monte Carlo* MCMC algorithm to simulate parameter densities and we employ the *deviance information criterion* DIC for determining the number of the structural changes. Section 2 introduces the AR(1)-GARCH(1,1) model and the Bayesian MCMC approach and Sect. 3 discusses the empirical results of four currencies: THB, PHP, IDR, and KRW. Conclusions are given in Sect. 4.

2 Model and Bayesian Inference

2.1 The Volatility Model

We assume a K-state Markov switching model where each component $k = 1, \ldots, K.$ is a GARCH model with an AR(1) disturbance. Each component is assumed to have an unconditional mean μ_k and AR(1) coefficient ϕ_k.

$$y_t = f_k(y_t) = \mu_k + \phi_k(y_{t-1} - \mu_k) + \varepsilon_{k,t}, \quad \varepsilon_{k,t} \sim \mathcal{N}(0, h_{k,t}), \tag{1}$$

The conditional variance $h_{k,t}$ is allowed to change through time by a GARCH(1,1) process:

$$h_{k,t} = \omega_k + \gamma_k h_{k,t-1} + \alpha_k \varepsilon^2_{k,t-1}, \tag{2}$$

The discrete random variables $\mathbf{s} = \{s_1, \ldots, s_t, \ldots, s_T\}$ are the state indicators at time t, and $s_t \in \{1, \ldots, k, \ldots, K\}$, follows a Markov process with transition matrix $\mathbf{\Pi}$ with K states

$$s_t \sim Markov(\mathbf{\Pi}), \tag{3}$$

and the elements of the probability π_{ij} of $\mathbf{\Pi}$ are given by

$$\pi_{ij} = P(s_t = j | s_t = i), \quad i = 1, \ldots, K, \quad j = 1, \ldots, K. \tag{4}$$

As the regime changes, the indicator s_t changes from 1 to K in ascending ordering. Therefore we define for $\mathbf{\Pi}$ a restricted (step-up) transition probability matrix following the approach of Chib (1998). Under these settings, the observation equation is a mixture model

$$y_t = \sum_{k=1}^{K} \mathbf{1}(s_t = k) f_k(y_t), \tag{5}$$

where $\mathbf{1}(\cdot)$ is a indicator variable for the event in the parenthesis. If the condition is true, then the indicator is 1, otherwise zero. The likelihood function is given by

$$\begin{aligned} L(\mathbf{y} \mid \mathbf{\Theta}) &= f(y_1 \mid \mathbf{\Theta}_1) \prod_{t=2}^{T} \sum_{k=1}^{K} f(y_t \mid \mathbf{y}_{t-1}, s_t, \mathbf{\Theta}) P(s_t = k \mid \mathbf{y}_{t-1}, \mathbf{\Theta}), \\ &= f(\varepsilon_{1,1} | \mathbf{\Theta}_1) \prod_{t=2}^{T} \sum_{k=1}^{K} f(\varepsilon_{k,t} | \mathbf{I}_{t-1}, \mathbf{\Theta}_k) P(s_t = k | \mathbf{I}_{t-1}, \mathbf{\Theta}) \end{aligned} \tag{6}$$

where $\mathbf{y} = (y_1, \ldots, y_T)'$, $\mathbf{y}_t = (y_1, \ldots, y_t)'$, $\mathbf{I}_t$ is the available information set at time t, and $\mathbf{\Theta} = \{\mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_K, \mathbf{\Pi}\}$, $\mathbf{\Theta}_k$ is the parameter vector associated with model k, namely $(\mu_k, \phi_k, \omega_k, \gamma_k, \alpha_k)'$.

Under these assumptions the likelihood function is,

$$f(\varepsilon_{1,1} \mid \mathbf{\Theta}_1) = \frac{1}{\sqrt{2\pi \frac{h_{1,1}}{1-\phi_1^2}}} \exp\left(-\frac{(y_1 - \mu_1)^2}{2\frac{h_{1,1}}{1-\phi_1^2}} \right), \tag{7}$$

$$f(\varepsilon_{k,t} \mid \mathbf{I}_{t-1}, \mathbf{\Theta}_k) = \frac{1}{\sqrt{2\pi h_{k,t}}} \exp\left(-\frac{(y_t - \mu_k - \phi_k(y_{t-1} - \mu_k))^2}{2h_{k,t}} \right). \tag{8}$$

We can evaluate the likelihood function using the GARCH densities as in Hamilton (1989) method independently only under Haas et al. (2003) formulations without any approximations for the GARCH model. This makes the likelihood of the models easy to be evaluated. Because of the non-linearity and the many parameters for large K, classical maximum likelihood methods requiring numerical optimizations are difficult to apply. In this situation, Bayesian MCMC methods yield faster parameter estimates without any optimizations.

2.2 The Prior and Posterior Distribution

Let $\mathbf{\Theta}$ be the parameter set of the K-state model and we assume that the prior for $\mathbf{\Theta} = \{\mathbf{\Theta}_1, \ldots, \mathbf{\Theta}_K, \mathbf{\Pi}\}$ is block-wise independent:

$$p(\mathbf{\Theta}) = p(\mu)p(\phi)p(\theta)\prod_{k=1}^{K-1} p(\pi_{kk}) \tag{9}$$

with $\mu = (\mu_1, \mu_2, \ldots, \mu_K)', \phi = (\phi_1, \phi_2, \ldots, \phi_K)', \theta = (\theta_1', \theta_2', \ldots, \theta_K')'$, $\theta_k = (\omega_k, \gamma_k, \alpha_k)', k = 1, \ldots, K$. For the vector of mean coefficients μ we assume

$$\mu \sim \mathcal{N}(\mu_{0,\mu}, \mathbf{\Sigma}_{0,\mu}), \quad \mu_{0,\mu} = \mathbf{0}_{K\times 1}, \mathbf{\Sigma}_{0,\mu} = 1{,}000 \times \mathbf{I}_{K\times K},$$

and for the AR(1) coefficients vector ϕ, we assume independent uniform prior for each element ϕ_k,

$$\phi_k \sim \mathcal{U}(-1, 1), \quad k = 1, \ldots, K,$$

To assure stationarity, the prior density is truncated to the interval $(-1, 1)$.

For the prior of the GARCH parameters, we assume a truncated normal density

$$\theta \sim \mathcal{N}(\mu_{0,\theta}, \mathbf{\Sigma}_{0,\theta}), \quad \mu_{0,\theta} = \mathbf{0}_{3K\times 1}, \mathbf{\Sigma}_{0,\theta} = 1{,}000 \times \mathbf{I}_{3K\times 3K}.$$

where the truncation is implied by imposing positive variances as a condition for the GARCH model and are given for each state $i = 1, \ldots k$ by

$$\omega_i, \gamma_i, \alpha_i > 0, \quad \text{and} \quad \gamma_i + \alpha_i < 1,$$

For the non-zero probabilities elements of the step-up transition matrix we use the beta distribution

$$\pi_{ii} \sim \mathcal{B}(a, b),$$

and following Chib (1998) we use the hyper-parameters $a = 9, b = 0.1$.

The posterior distribution is – by Bayes's theorem – proportional to multiplying (9) and (6)

$$f(\mathbf{\Theta} \mid \mathbf{y}) \propto L(\mathbf{y} \mid \mathbf{\Theta})p(\mathbf{\Theta}). \tag{10}$$

Table 2 Model choice by DIC for five regimes (minimum DIC in bold)

	THB	PHP	IDR	KRW
k = 1	3864.64	2374.26	8079.09	3899.39
k = 2	3682.10	1871.54	7963.10	3822.86
k = 3	3693.61	**1855.72**	**7830.75**	**3808.48**
k = 4	**3589.26**	2006.95	7986.91	5698.24
k = 5	4022.80	2549.86	7984.27	4470.55

2.3 Gibbs Sampling

This section develops a MCMC algorithm for the K-state AR(1)-GARCH(1,1) model and lists all necessary full conditional distributions for the posterior in (10). The MCMC sampling scheme with Metropolis-Hastings (MH) steps comprises:

0 initialize $\boldsymbol{\Theta}^{(0)}$,
1 draw π_{ii} from a beta distribution (see Kim and Nelson 1999),
2 draw θ using a random walk MH algorithm (see Holloway et al. 2002),
3 draw ϕ using the MH algorithm (see Chib and Greenberg 1995),
4 draw μ from a normal distribution,
5 draw $\mathbf{s}$ from a Bernoulli distribution (see Kim and Nelson 1999).

we iterate step 1–5 for $G = 50{,}000$ times and we discard 10,000 iterations as burn-in.

3 Empirical Analysis

We consider the daily log returns of four Asian currencies against the USD from Jan. 3, 1995 to June 30, 2008: Thailand's Baht THB, Philippine Peso PHP, Indonesia Rupiah IDR, and S. Korea KRW (with sample sizes 3,387, 3,152, 3,140, and 3,390).

3.1 Model Choice

To determine the adequate number of regimes K, we calculate the dispersion information criterion DIC suggested by Spiegelhalter et al. (2002). Table 2 lists the DIC's for up to the $K = 1, \ldots, 5$ regimes. Only for the Thai THB we find three structural changes (as the minimum DIC for K is 4) between 1995 and 2008, while for the other three currencies

To estimate the exact date of the structural change, we have computed the posterior probability of the states s_t using $s_t^{(g)}$ from the MCMC sample:

$$\hat{P}(s_t = i) = G^{-1} \sum_{g=1}^{G} \mathbf{1}(s_t^{(g)} = i), \quad i = 1, \ldots, K.$$

In the Markov switching model, the state s_t is estimated through the largest posterior probability $\hat{P}(s_t = i)$ and the estimated regime changes are shown in Table 3.

The time series of estimated posterior volatilities and the probability of a break point for the four countries are shown in Figs. 1–4.

3.2 Thailand

Thailand's currency changed three times the volatility regime and Table 3 shows that the first break point occurred on May 15, 1997. Recall from Table 1 that the Asia crisis started by serious attacks of hedge funds on May 14 and 15 against the Thai currency, so the break point marks exactly the beginning of the Asian crisis. After fruitless defenses the authorities had to change the currency system on July 2, 6 weeks after the attacks began, thus the estimated structural change in volatilities is exactly in line with financial history.

The second break point occurred on Sep. 29, 1998, five quarters after the regime switch, and marked the end of the high volatilities, and about 1 month after a new agreement with the IMF was found.[1]

Table 3 Estimated dates of the break points

Break Points	1st	2nd	3rd
Thailand	5/15, 1997	9/29 1998	12/11, 2006
Philippines	5/7, 1996	7/4, 1997	N/A
Indonesia	7/15, 1997	10/9, 2001	N/A
South Korea	1/30, 1996	1/23, 1998	N/A

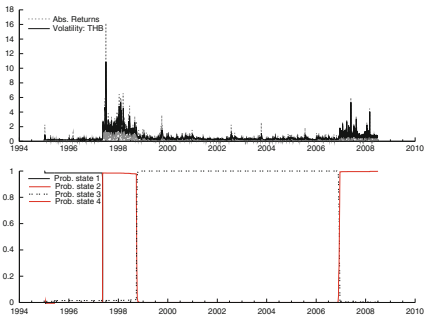


Fig. 1 Thailand THB

[1] On Oct. 3, 1998, Japan declared "A New Initiative to Overcome the Asian Currency Crisis" (or New Miyazawa Initiative) to help Asian economies and the stability of financial markets. Japan

Fig. 2 Philippine PHP

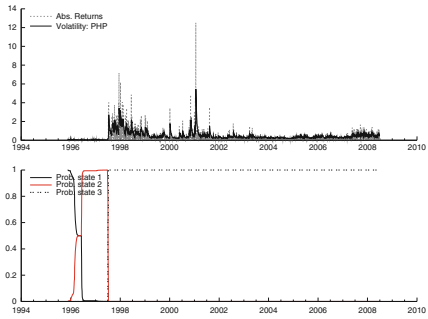


Fig. 3 Indonesian IDR

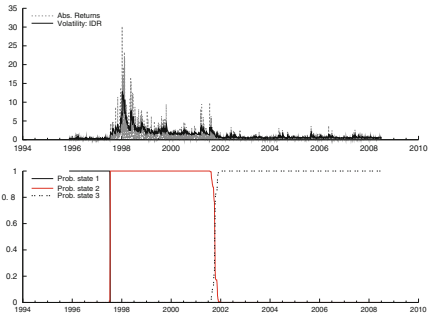
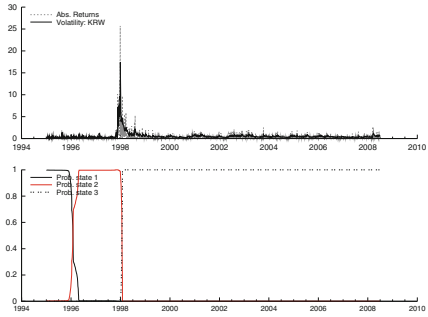


Fig. 4 South Korean KRW



The third break point is estimated for Dec. 11, 2006, after which the volatility increased almost to the level of the Asia crises. This increase in volatility was caused by a coup d'etat, which took place on Tuesday 19 Sep. 2006, when the Thai Army toppled the elected government of Prime Minister Thaksin Shinawatra. The subsequent instability of the new government of Thailand made the currency more volatile.

provided a package of US$30 billion for the economic recovery in Asia. It seems that the funding program has helped to stabilize the markets.

3.3 The Philippines

The Philippine currency dynamics changed two times according to the estimates of posterior probability in Fig. 2. The first break point in the currency volatilities is July 11, 1996, 1 year before the Asia crisis and from the estimated volatilities we see that the size and duration was negligible compared with the Asia crises. Table 3 shows that the second break point was on July 4, 1997, 2 days after the Thai authorities had to change their currency system. It shows that the currency has become volatile before the currency system changed and after the Philippine central bank raised interest rates by 3.75% points in defense of the peso in spring 1997. After the second structural change point, the high PHP volatilities have continued for a long time and the stabilization worked rather gradually.

3.4 Indonesia

Similar to Thailand, the first volatility break point occurred on July 15, 1997, right 1 month before the date of change in the currency system on Aug. 14, 1997, and the volatility of the IDR went up. The next change point is estimated for Oct. 9th, 2001, as can be seen in Fig. 3 and Table 3. President Abdurrahman Wahid was discharged on July 23th 2001, as he broke with the IMF. The next president Diah Permata Megawati Setiawati Sukarnoputri, the daughter of the former president Sukarno, restored the relationship between Indonesia and the IMF. After the event, the IMF resumed the financial support for Indonesia on Sep. 10, 2001, and the long 4-year period of high volatilities came to an end. This shows that the restart of IMF funding policy stabilized the Indonesia IDR despite the coincidence of the 9/11 attacks which had no effects on the Asian currencies.

3.5 South Korea

South Korea's currency changed two times the regime according to the estimates shown in Fig. 4. The first change occurred on Jan. 30, 1996, 2 years before the second one, during the peg regime and was quite small. This first sign of trouble in Korea became evident, when the current account deficit widened from 2% of GNP in 1995 to 5% in 1996. The subsequent change in the currency occurred on Dec. 16, 1997, and was the latest of the four countries considered in this study. This came after some serious drops in the stock markets at the end of the year together with a downgrading from A1 to B2 in Moody's credit rating.

Table 3 shows that the second break point is on January 23, 1998, 1 month after the currency system has changed. Thus the currency became shortly volatile after the change from peg to floating. This raises an interesting issue: Why were there no speculative attacks on the KRW and is this the reason of a delayed volatility response in the currency?

4 Conclusions

This paper has analyzed the structural changes in the volatilities of Asian currencies over the 1995–2008 decade covering the Asia and the "dot.com" crises and the slump following the 9/11 attacks. We find that strong volatility changes had occurred for the Thai THB and Indonesian rupiah caused by the Asian crisis.

In the introduction we asked: Were the changes in the four currency systems accompanied by similar patterns in the volatility structure? The answer is rather no, despite the fact that the four countries changed from peg to float in the second half of 1997. Vulnerability, duration and the response dates to currency attacks seem to be quite different and to depend on the underlying strength of the economies. Occasionally we find similar patterns of currency changes like the one for Thailand and Indonesia. Furthermore, some regime shifts are strongly related to the internal politics of the countries, like discharge of presidents or coup d'etat. And we find that quite intensive influence by the IMF or other countries can be an effective way to stabilize the volatility of the currencies.

We find that that the effects of the Asia crises are quite divers if we only concentrate on currency fluctuations. More can be learned if we take into account the effects of the stock markets and interest rates, or growth and deficits. But the modeling complexity will not decrease since relationships between countries will not become easier in times of a crisis. But our modeling approach shows that regime shift models can resolve some of the complexities that are triggered by crises developments.

References

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31*, 307–327.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics, 86*, 221–241.

Chib, S., & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *American Statistician, 49*, 327–335.

Haas, M., Mittnik, S., & Paolella, M. S. (2003). A new approach to Markov-Switching GARCH Models. *Journal of Financial Econometrics*, 95–530.

Hamilton, J. D. (1989). A new approach to the econometric analysis of nonstationary time series and the business cycle. *Econometrica, 57*, 357–384.

Holloway, G., Shankar, B., & Rahman, S. (2002). Bayesian spatial probit estimation: A primer and and application to HYV rice adoption. *Agricultural Economics, 27*, 383–402.

Kim, C., & Nelson, C. R. (1999) *State-space models with markov switching.* Cambridge: The MIT Press.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society*. Series B 64, 583–639.

The Pricing of Risky Securities in a Fuzzy Least Square Regression Model

Francesco Campobasso, Annarita Fanizzi, and Massimo Bilancia

Abstract This work aims to estimate the relationship between the expected return of a financial investment and its risk by means of a fuzzy version of the Capital Asset Pricing Model (CAPM). The expected return is usually computed as a function both of the rate of a risk-free security, that represents the time value of money, and of a premium that compensates investors for taking on an additional risk in the market. Actually we estimate the parameters of a simple regression model, where the dependent variable consists in the percentage change in prices of a surveyed (stable or volatile) stock and the independent variable consists in the percentage change in market indexes. As both changes in closure prices only partially represent the actual trend in returns, we use a range of observed values for each price; this allows us to estimate the sensitiveness of the stock to risk by means of the so called Fuzzy Least Square Regression. The corresponding estimates are compared with the ones obtained by means of the Ordinary Least Square Regression.

1 The Capital Asset Pricing Model

According to the Capital Asset Pricing Model (Black et al. 1972), the expected return of a stock depends both on the rate of a risk-free security, that represents the time value of money, and on a premium for taking on an additional risk. In formal terms the aforesaid relationship can be written as

$$E\left(r_{it} - r_f\right) = \beta_i \left[E(r_{mt}) - r_f\right] \tag{1}$$

where

- r_{it} is the risky return of the i-th stock at time t, which equals the percentage change in its prices between times $t-1$ and t;

M. Bilancia (✉)
Dipartimento di Scienze Statistiche "Carlo Cecchi", Università degli Studi di Bari, Italy
e-mail: mabil@dss.uniba.it

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_70,

- r_{mt} is th risky return of the market portfolio at time t, which equals the percentage change in levels of the market index between times $t-1$ and t;
- r_f is the risky free return, i.e. the return of a short run Treasury Bond.

Let us suppose that r_f is constant, as investors know the risk-free return when they choose their portfolios. After defining the unexpected return of the i-th stock at time t as

$$u_{it} = r_f - E(r_{it}) \tag{2}$$

and the unexpected return of the market portfolio at time t as

$$u_{mt} = r_f - E(r_{mt}) \tag{3}$$

the model (1) can take the following form

$$r_{it} - r_f = \beta(r_{mt} - r_f) + \varepsilon_{it} \tag{4}$$

where $\varepsilon_{it} = u_{it} - \beta_i u_{mt}$ satisfies the classic Gauss-Markov assumptions.

Noting that only r_f affects the intercept in the linear relationship between r_{it} and r_{mt}, we can specify model (1) as

$$r_{it} = \alpha_i + \beta_i r_{mt} + \varepsilon_{it} \tag{5}$$

where $\alpha_i = r_f + \beta_i r_f$ represents such an intercept. The regression coefficient β_i measures the way in which the price of the i-th stock varies according to the market. We can easily obtain its estimate by means of Ordinary Least Squares (OLS) method. Depending on whether $\hat{\beta}_i = 1$, $\hat{\beta}_i > 1$ or $\hat{\beta}_i < 1$, the premium for the risk of the i-th stock may equal, exceed or underlie the premium for the market risk.

The estimate of the regression coefficient β_i may be improper, as the percentage changes in closure prices (both of the i-th stock and of the market index) only partially represent the actual trend of the corresponding returns. We propose to use a range of various percentage changes varying between the following two ratios (both decreased by 1): the ratio of the highest price of the i-th stock (market index) at time t to the lowest price at time $t-1$, and the ratio of the lowest price of the i-th stock (market index) at time t to the highest price at time $t-1$. Accordingly, it is worth noting that

$$\frac{\text{lowest price}(t)}{\text{highest price}(t-1)} - 1 \leq \frac{\text{closure price}(t)}{\text{closure price}(t-1)} - 1 \leq \frac{\text{highest price}(t)}{\text{lowest price}(t-1)} - 1 \tag{6}$$

The more the values of the range roll away from the percentage change in closure prices, the less they are likely. This feature of the suggested values allows us to introduce triangular fuzzy numbers in the Capital Asset Pricing Model.

2 A Fuzzy Least Squares Regression Approach

Fuzzy techniques can be used to fit data into a regression model, where deviations between dependent variable and its model are connected with the uncertain nature whether of independent variables or of their coefficients. There are two lines of investigation developed in literature in this regard.

The first one (Tanaka et al. 1989) considers as nuanced the relationship between observed variables and, therefore, uses regression coefficients of fuzzy type; these are estimated minimizing through linear programming their total spread, representing uncertainty of the model (Fuzzy Possibilistic Regression).

The second approach (Diamond 1988) is similar to traditional one, because regression coefficients are estimated minimizing the distance between observed fuzzy values and corresponding theoretical values (Fuzzy Least Square Regression). This work follows on from this approach, whose main merit consists in introducing a metric onto the space of fuzzy numbers, although necessarily triangular.

A fuzzy set $\widetilde{X}$ for the statistical variable X is defined as the set of pairs $\left\{x_i, \mu_{\widetilde{X}}\right\}$, where x_i represents any possible value of X and $\mu_{\widetilde{X}}(x_i) : X \to [0, 1]$ expresses its membership degree to $\widetilde{X}$. Finally a fuzzy number is any fuzzy set which is at once normal and convex.

A triangular fuzzy number (Zimmermann 1991) $\widetilde{X} = (x, x^L, x^R)_T$ for the variable X is characterized by a membership function $\mu_{\widetilde{X}} : X \to [0, 1]$, see the one represented in Fig. 1, that expresses the membership degree of any possible value of X to $\widetilde{X}$. The accumulation value x is considered the centre of the fuzzy number, while $x - x^L$ and $x^R - x$ are respectively considered as the left spread and the right spread. Note that x belongs to $\widetilde{X}$ with the highest degree, while the other values included between the extremes x^L and x^R belong to $\widetilde{X}$ with a gradually lower degree.

A metric onto the space onto the space of triangular fuzzy numbers was introduced in Diamond (1988), according to which the distance between and $\widetilde{X}$ and $\widetilde{Y}$ is

$$d(\widetilde{X}, \widetilde{Y})^2 = (x - y)^2 + (x^L - y^L)^2 + (x^R - y^R)^2 \tag{7}$$

On the ground of these considerations we analyzed the performance of financial investments through the Fuzzy Least Square Regression, that covers the case of a dependent fuzzy variable $\widetilde{Y}_t$ (the percentage change in the prices of a specific

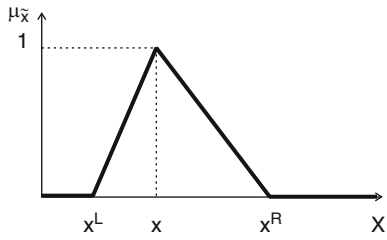


Fig. 1 Representation of a triangular fuzzy number

stock) in terms of an independent fuzzy variable $\widetilde{X}_t$ (the percentage change in the prices of the market index). In particular, if we observe $\widetilde{Y}_t = (y_t, y_t^L, y_t^R)_T$ and $\widetilde{X}_t = (x_t, x_t^L, x_t^R)_T$ at time t, the regression model becomes

$$\widetilde{Y}_t = \alpha + \beta \widetilde{X}_t + \varepsilon_t \tag{8}$$

The expression of the estimated parameters of the model is derived from minimizing the sum $\sum_{t=1}^{n} d(\alpha + \beta \widetilde{X}_t, \widetilde{Y}_t)^2$ of the squared distances between theoretical and empirical values of the fuzzy dependent variable with respect to α and β.

It was proved in Diamond (1988) that the Fuzzy Least Squares optimization problem has, under mild conditions, the following unique solutions

$$\hat{\alpha} = \frac{1}{3}\left[(\overline{y} + \overline{y}^L + \overline{y}^R) - \hat{\beta}(\overline{x} + \overline{x}^L + \overline{x}^R)\right] \tag{9}$$

$$\hat{\beta} = \frac{3\sum_{i=1}^{n}\left[(x_t^R)(y_t^R) + (x_t^L)(y_t^L) + x_t y_t\right] - n(\overline{y} + \overline{y}^L + \overline{y}^R)(\overline{x} + \overline{x}^L + \overline{x}^R)}{3\sum_{i=1}^{n}\left[x_t^2 + (x_t^R)^2 + (x_t^L)^2\right] - n(\overline{x} + \overline{x}^L + \overline{x}^R)} \tag{10}$$

where $\overline{y}$, $\overline{y}^R$ and $\overline{y}^L$ ($\overline{x}$, $\overline{x}^R$ and $\overline{x}^L$) represent respectively the average accumulation value and the average extreme values, both on the right and on the left, of the fuzzy dependent (independent) variable.

The expressions of the parameters $\hat{\alpha}$ and $\hat{\beta}$ are similar to those based on Ordinary Least Squares, differing only for the adopted metrics. It should be noted that the total of theoretical accumulation and extreme values, both on the right and on the left, of the dependent variable coincides with the same amount of empirical values, but the average estimated values of the dependent variable does not match the average empirical values, unlike in classical regression. In particular the estimated regression coefficient $\hat{\beta}$ is still equal to the ratio of the covariance between the two observed variables on the variance of the independent variable, both expressed in fuzzy terms.

As the same author says, the computational simplicity of the estimation procedure offsets the severity of its restriction to a certain type of fuzzy numbers (triangular ones).

In order to evaluate how the Fuzzy Least Square Regression fits the data, we propose a fuzzy version of R^2 index, that can be called Fuzzy Fit Index (FFI). Its expression can be still obtained as the ratio between regression deviance and total deviance, clearly expressed in accordance with the introduced metrics

$$\text{FFI} = \frac{\sum_{i=1}^{n} d(\widetilde{Y}_t^{\star}, \overline{Y}^{\star})^2}{\sum_{i=1}^{n} d(\widetilde{Y}_t, \overline{Y})^2} \tag{11}$$

where $\overline{Y}^{\star}$ and $\overline{Y}$ represent the averages respectively of the theoretical and of the observed values of the dependent variable, while $\widetilde{Y}_t^{\star}$ are the fitted values on the Fuzzy Regression line. The closer this index is to one, the better the model fits the observed data (Campobasso et al. 2008). Note that FFI not necessarily

increases when the number of explanatory variables included in the model growths; at variance with classical regression, the average estimated values of the dependent variable does not match the average empirical values, as we stated above.

3 Case Study

The case study refers to the Italian financial market in the period between January 2003 and January 2009. In order to test the ability of the model to forecast quotations both of a stable stock and of a volatile one on the basis of the fluctuations in the market, we consider the percentage change in prices of the MIBTEL index as the independent variable, while the percentage change in prices of the TISCALI stock first and then of the ENEL stock as the dependent variable (Fig. 2). Specifically, in order to analyze if the frequency of observations affects the estimates in the model, we consider daily, weekly and monthly closures.

In the case of a volatile stock (Table 1), both the fuzzy and the classic approach determine a regression coefficient β greater than 1, so that the premium for the stock risk appears anyway greater than the premium for the market risk. Specifically the

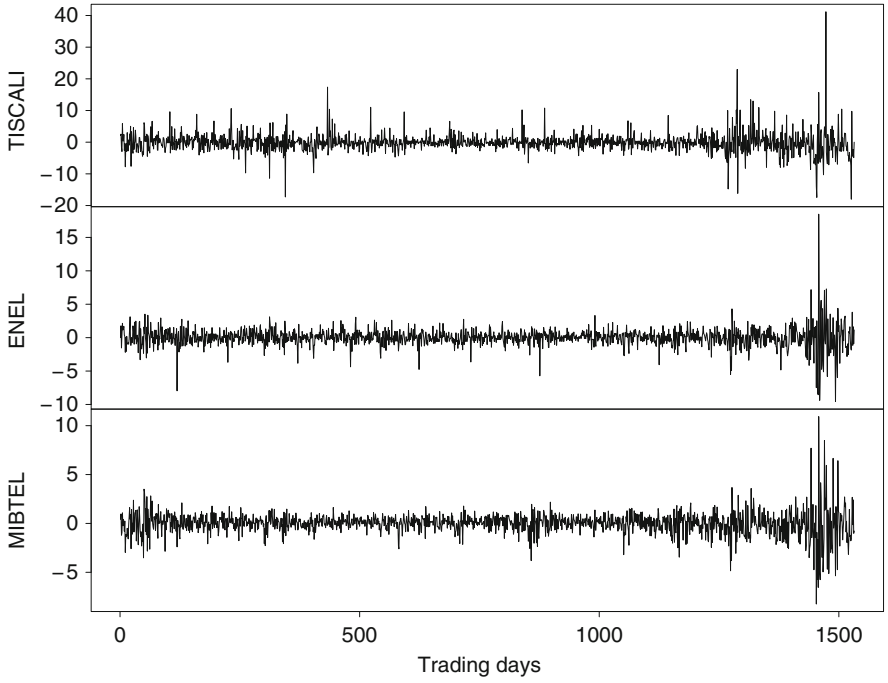


Fig. 2 Percentage change in daily prices between January 2nd, 2003 and January 31th, 2009. Two stocks, TISCALI and ENEL, as dependent variable, and the composite MIBTEL index, as independent variable, have been considered

Table 1 The estimates for the TISCALI stock

	Daily prices		Weekly prices		Monthly prices	
Estimates	**Fuzzy**	**Classic**	**Fuzzy**	**Classic**	**Fuzzy**	**Classic**
Intercept	0.00	–0.11	0.00	–0.51	–0.02	–1.91
Mibtel	2.01	1.09	2.15	1.24	2.30	1.93
FFI	0.52	–	0.65	–	0.64	–
$\mathbf{R^2}$	–	0.17	–	0.24	–	0.35

Table 2 The estimates for the ENEL stock

	Daily prices		Weekly prices		Monthly prices	
Estimates	**Fuzzy**	**Classic**	**Fuzzy**	**Classic**	**Fuzzy**	**Classic**
Intercept	0.00	0.01	0.00	0.03	0.00	–0.06
Mibtel	1.10	0.88	1.03	0.82	1.00	0.71
FFI	0.84	–	0.85	–	0.94	–
$\mathbf{R^2}$	–	0.51	–	0.50	–	0.47

fuzzy approach determines a definitely greater β and fits the data much better than the classic one. Moreover the sensitiveness of the stock to market fluctuations grows from daily to monthly lists (maybe because the stock does not react immediately to any change in the market).

In case of a stable stock (Table 2), the fuzzy approach determines a regression coefficient β slightly greater than 1, whereas the classic approach determines a β smaller than 1; therefore the performance of the stock results more or less rewarding than the market one depending on the selected estimation procedure. However the fuzzy model fits the data in a better way. Moreover the sensitiveness of the stock to market fluctuations decreases from daily to monthly lists (maybe because its reactions tend to smooth over time). In both cases (TISCALI and ENEL) the intercept of the fuzzy model almost equals zero, because the estimation procedure takes into account the fluctuations of the risk-free return (generally minimal and assumed constant) during an economic growth or regression.

The goodness of fit of our proposal is also tested using available data beyond the estimation period. By taking into account daily and weekly quotations of the considered stocks from February 2009 to May 2009, we compare the obtained fuzzy spreads with the classic confidence intervals in order to verify whether empirical observations are included in both such ranges. Monthly quotations are omitted, as their fluctuations are unduly smooth. It is possible to observe that the percentage change in prices of the stable stock (ENEL) always lies between fuzzy extremes, whereas it does not in confidence intervals (Fig. 3); this happens in the case both of daily and of weekly quotations (especially in the first one). The same is true for weekly and, albeit to a lesser extent, for daily quotations of the volatile stock (TISCALI), that cannot be reliably estimated by means of classic methods (Fig. 4).

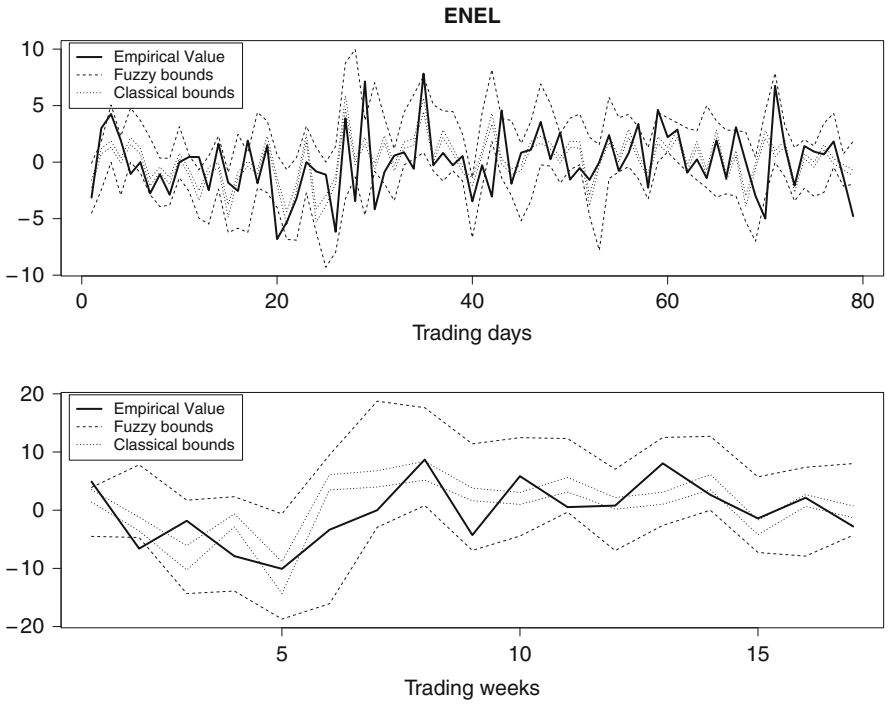


Fig. 3 Percentage change and corresponding confidence bounds in daily and weekly prices of ENEL stock between February 1st, 2009 and May 31th, 2009

4 Final Remarks

A fuzzy version of the Capital Asset Pricing Model allows to better estimate the sensitivity of a stock to market fluctuations, because it uses additional information on the observed trends. A simple application demonstrates how such a version is better suited to a fast moving reality: the stronger the variability of a stock is, the more suitable representation of its expected return is provided. In this case, in fact, the typical hypothesis of the Capital Asset Pricing Model (that the risk-free return in a market is constant over time) can not occur in practice and yet its fuzzy version blurs any changes of such a return over time.

The Fuzzy Fit Index represents the portion of variance of the stock return explained by the market index and provides a measure of the systematic risk (impossible to remove by means of diversification); the remaining part of the variance represents the specific risk due to the volatility of the stock (that can be managed by means of diversification). Such an index arises if the model includes stable stocks (which are characterized by a low variability in prices over time) and however if the frequency of the observed data decreases, maybe because the fluctuations in prices become more regular. In particular the expected premium for the risk of a volatile stock exceeds the expected premium for the market risk; moreover the sensitiveness

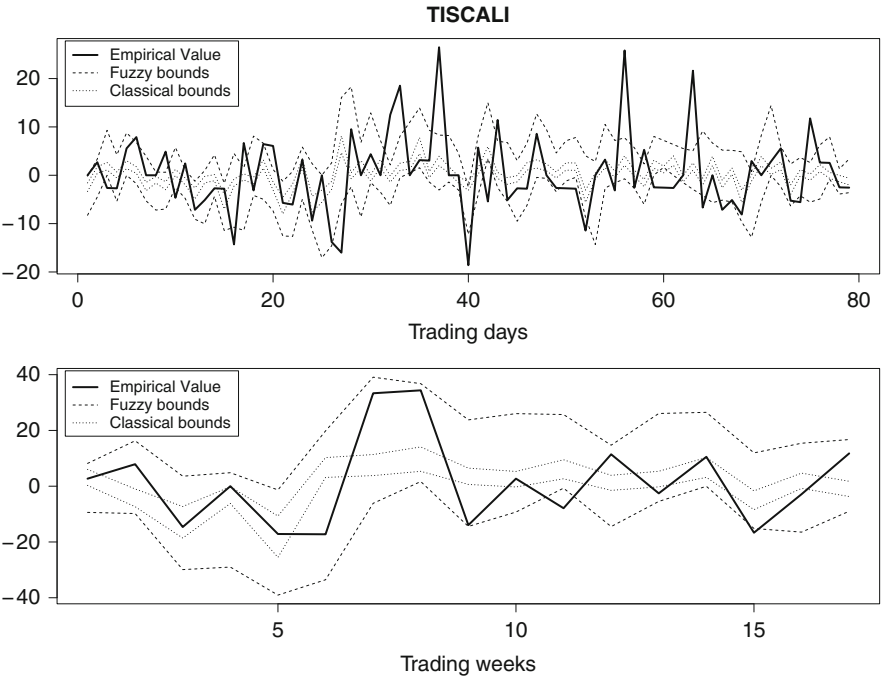


Fig. 4 Percentage change and corresponding confidence bounds in daily and weekly prices of TISCALI stock between February 1st, 2009 and May 31th, 2009

to market fluctuations of such a stock grows when the frequency of the observed data decreases. On the contrary the expected premium for the risk of a stable stock almost equals the expected premium for the market risk; moreover the sensitiveness of such a stock tends to smooth when the frequency of the observed data decreases.

Acknowledgements Francesco Campobasso conceived the study, wrote Sects. 3 and 4, and revised the draft manuscript. Annarita Fanizzi wrote Sect. 2. Massimo Bilancia wrote Sect. 1. The authors read and approved the final manuscript.

References

Black, F., Jensen, M., & Scholes, M. (1972). The capital asset pricing model: some empirical test. In M. Jensen (Ed.), *Studies in the theory of capital markets*, New York: Praeger Publishers Inc.

Campobasso, F., Fanizzi, A., & Tarantini, M. (2008). Una generalizzazione multivariata della fuzzy least square regression. *Annali del Dipartimento di Scienze Statisiche "Carlo Cecchi" – Università di Bari*, 229–230.

Diamond, P. (1988). Fuzzy least squares regression. *Information Sciences, 46*, 151–157.

Tanaka, H., Hayashi, I., & Watada, J. (1989). Possibilistic linear regression analysis for fuzzy data. *European Journal of Operational Research, 40*, 389–396.

Zimmermann, H. J. (1991). *Fuzzy set theory*. Dordrecht, The Netherlands: Kluwer Academic Publishing.

Classification of the Indo-European Languages Using a Phylogenetic Network Approach

Alix Boc, Anna Maria Di Sciullo, and Vladimir Makarenkov

Abstract Discovering the origin of the Indo-European (IE) language family is one of the most intensively studied problems in historical linguistics. Gray and Atkinson (2003) inferred a phylogenetic tree (i.e., additive tree or X-tree, Barthelémy and Guénoche 1991) of the IE family, using bayesian inference and rate-smoothing algorithms, based on the 87 Indo-European language data set collected by Dyen et al. (1997). When conducting their classification study, Gray and Atkinson assumed that the evolution of languages was strictly divergent and the frequency of word borrowing (i.e., horizontal transmission of individual words) was very low. As consequence, their results suggested a predominantly tree-like pattern of the IE language evolution. In our opinion, only a network model can adequately represent the evolution of the IE languages. We propose to apply a method of horizontal gene transfer (HGT) detection (Makarenkov et al. 2006) to reconstruct a phylogenetic network depicting the evolution of the IE language family.

1 Introduction

A number of curious parallels between the processes of historical linguistics and species evolution have been observed (Atkinson and Gray 2005; Gray and Atkinson 2003; Rexová et al. 2003). The evolutionary biologists and historical linguists often look for answering similar questions and face similar problems (Atkinson and Gray 2005). Recently, the theory and methodology of the two fields have evolved in remarkably similar ways. A number of important studies have considered the applications of phylogenetic methods to process language data (e.g., Atkinson and Gray 2005; Gray and Atkinson 2003; Rexová et al. 2003). For instance, one of the most intensively studied topics is the evolution of the Indo-European (IE) language family (Diamond and Bellwood 2003). Gray and Atkinson (2003) inferred a consensus

A. Boc (✉)
Université du Québec à Montréal, Case postale 8888, succursale Centre-ville
Montréal (Québec) H3C 3P8 Canada
e-mail: boc.alix@courrier.uqam.ca

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_71,

phylogenetic tree of the IE language family using maximal likelihood models of lexical evolution, bayesian inference and rate-smoothing algorithms; the 87 Indo-European language data set collected by Dyen et al. (1997) was analyzed in Gray and Atkinson (2003). On the other hand, Rexová et al. (2003) also reconstructed a phylogeny of the IE languages when applying a cladistic methodology to study the same lexicostatistical data set (Dyen et al. 1997). The results obtained in Rexová et al. (2003) were very similar to those found in Gray and Atkinson (2003). However, to reconstruct their phylogenies Gray and Atkinson, as well as Rexová et al., were constrained to assume that the evolution of languages was strictly divergent, each language was transmitted as a whole, and the frequency of borrowing (i.e., horizontal transmission of individual words) between languages was low. As consequence, the obtained results suggested a predominantly tree-like pattern of the IE language evolution with little borrowing of individual words.

In our opinion, only a phylogenetic network can adequately represent the evolution of this language family. A network model can incorporate the borrowing and homoplasy (i.e., evolutionary convergence) processes that influenced the evolution of the Indo-European languages. For example, although English is a Germanic language, it has borrowed around 50% of its total lexicon from French and Latin (Pagel 2000).

We propose to apply the methods of horizontal gene transfer (HGT) detection, which are becoming very popular among molecular biologists, in order to reconstruct the evolutionary network of the IE language family. The most frequent *horizontal word transfers*, representing borrowing events, will be added to the phylogenetic tree inferred by Gray and Atkinson (Fig. 1 in Gray and Atkinson 2003) to represent the most important word exchanges which occurred during the evolution of the IE languages. In particular, a HGT detection algorithm (Makarenkov et al. 2006) will be applied to build the evolutionary network of the IE languages.

In this article, we first outline the data in hand and then describe the new features of the HGT detection algorithm used to identify the word borrowing events. In the Results and Discussion section, we present the obtained results for the 12 most important groups of the IE languages and report the words borrowing statistics. The most important word exchanges characterizing the evolution of this language family will be brought to light and discussed.

2 Description of the Dyen Database

The database developed by Dyen et al. (1997) includes the 200 words of the Swadesh list (Swadesh 1952). The Swadesh list is one of several lists of vocabulary with basic meanings, developed by Morris Swadesh in the 1940–50s (Swadesh 1952), which is widely used in lexicostatistics (quantitative language relatedness assessment) and glottochronology (language divergence dating). Dyen et al. (1997) built a database that provides cognation data among 95 Indo-European speech varieties. For each word meaning in the list of 200 basic meanings (chosen by Swadesh

in 1952), the database contains the forms (e.g., words) used in the 95 speech varieties and the cognation decisions among the speech varieties made by Isidore Dyen in the 1960s. For each meaning, the forms were examined and cognation judgments were made (Dyen et al. 1997). The cognation judgments were made only between forms having the same meaning. The cognation judgments were recorded in classes of forms such that the forms in each class were "cognate" or "doubtfully cognate" with each other. Two forms, in two different speech varieties, were identified as "cognate" if within both of the varieties they had an unbroken history of descent from a common ancestral form. For example, since the English word FRUIT and French word FRUIT are known to be related by borrowing, they have been assigned different Cognate Classification Numbers (CCN) in the Dyen database (Dyen et al. 1997). Forms believed to be related by borrowing or by accidental similarity were thus not treated as cognate. In a small number of cases it was difficult to distinguish cognates from borrowings or accidental similarities; in this case the forms were classified as "doubtfully cognate" (Dyen et al. 1997). The cognate content information was used by Gray and Atkinson (2003) to reconstruct the evolutionary tree of IE languages. In our study, we also subdivided the 200 words of the Swadesh list into two broad categories: lexical (nouns and verbs; 138 words in total) and functional (adjectives + pronouns, conjunctions and determiners; 62 words in total) in order to see whether the rate of borrowing differs for these two broad categories.

3 Materials and Methods

In this section, we describe the new features of the HGT detection algorithm Makarenkov et al. (2006), applied here in a biolinguistics context, to infer a phylogenetic network of the IE languages family. When applied in a biological context, this algorithm identifies horizontal gene transfers (HGT) of a given gene for a given set of species thus reconciliating the species and gene phylogenetic trees. At each step of the reconciliation process, a HGT event is inferred. In this study, we draw a parallel between the HGT detection and the word borrowing detection processes. In our model, the IE languages tree (Fig. 1 and Fig. 4 in Gray and Atkinson 2003) plays the role of the species tree and the word tree, representing the evolution of a given word (a given translation in all 87 considered languages), plays the role of the gene tree. The algorithmic procedure includes the three main steps, which are as follows:

Step 1. Let L be the rooted tree of 87 IE languages inferred by Gray and Atkinson (2003). Figure 1 shows a representation of this tree by groups (the group content is reported on the right). We also considered the 200 words of the Swadesh list (Swadesh 1952) and their translations into 87 IE languages (Dyen et al. 1997). For each word of this list, we computed a distance matrix, $\mathbf{D}_i$ (87×87), $i = 1, \ldots, 200$, between its translations using a normalized Levenshtein distance [(1), Levenshtein 1966].

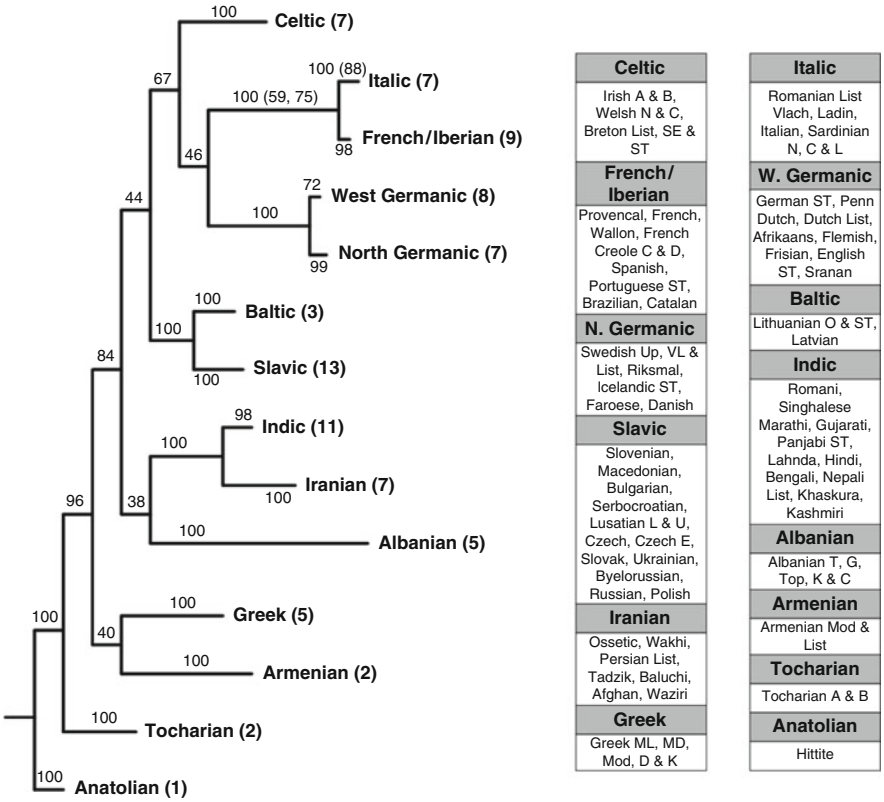


Fig. 1 Gray and Atkinson (Fig. 1 in Gray and Atkinson 2003) IE language evolutionary tree for 14 main language groups. The group content is indicated on the right. The numbers on the tree branches are the tree bootstrap scores; the number of languages for each group is indicated between *parentheses*

$$d(i,j) = \frac{Levenshtein\_distance(i,j)}{length(i) + length(j)}. \tag{1}$$

For each such matrix, we inferred the word phylogenetic tree W_i, using the Neighbor Joining method (Saitou and Nei 1987). Figure 2 shows the Robinson and Foulds (RF) topological distance (Robinson and Foulds 1981) (normalized by its maximal value of $2n - 6$ for two binary trees with n leaves) between each of the 200 word trees W_i and the language tree L. The average value of the normalized RF distance was 82%. Such a high value suggests an important overall discrepancy between the language tree L and the word trees W_i ($i = 1, \ldots, 200$).

Step 2. We applied the HGT detection algorithm (Makarenkov et al. 2006) to infer the word borrowing events, considering, in turn, the language tree L and each of the 200 word trees W_i. Therefore, 200 different scenarios of tree reconciliation were computed. As the Dyen database (Dyen et al. 1997) did not comprise any translation

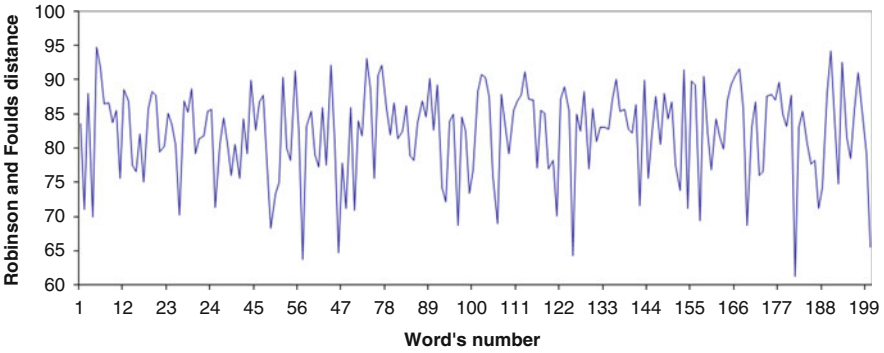


Fig. 2 Normalized Robinson and Foulds topological distance (Robinson and Foulds 1981) between each of the word trees and the IE language tree in Fig. 1

for the Hittite and Tocharian languages, belonging to the Anatolian and Tocharian groups respectively, these languages were not considered in our analysis.

Step 3. We combined all results from the obtained transfer scenarios to compute the word borrowing statistics. Intra-group transfers were ruled out in our computations because of the high risk of accidental similarity among the words from the same language group. First, we assessed the total numbers of transfers (i.e., number of word borrowings) between each pair of groups, and then the percentages of words affected by these transfers in each group. The 10 most important transfers were mapped into the IE language tree (see Figs. 3–5). These computations were carried out, first, for all 200 words and then, separately, for the words from the lexical and functional categories.

4 Results and Discussion

Figure 3 shows the total numbers of borrowed words found for each pair of groups. The 10 most active transfers are highlighted in dark grey. These transfers have been mapped into the IE language tree (Fig. 5a). If the geographical proximity can explain most of the frequent exchanges (e.g., between the West and North Germanic groups), some of them occur between the groups located far away from each other in the tree (e.g., between the Celtic and Indic, or Iranian and Celtic groups).

We can also observe a number of very active exchanges between the cluster combining the Indic and Iranian groups, and that combining the Celtic, Italic, French/Iberian, West/North Germanic and Slavic groups. These results suggest that despite the fact that the Iranian and Celtic groups are located far away from each other in the phylogenetic tree (Fig. 1), there is a strong relationship between them.

Figure 4 reports the percentages of words of a given group affected by transfers originating from other groups. Similarly to the results reported in Fig. 3, the highest values were found for the neighbor groups. One can also notice that the

	Celtic	Italic	French Iberian	W. Germanic	N. Germanic	Baltic	Slavic	Indic	Iranian	Albanian	Greek	Armenian
Celtic	-	53	82	88	58	16	68	89	83	52	30	37
Italic	54	-	357	29	24	10	45	49	32	33	18	1
French/Iberian	33	261	-	34	17	4	17	46	36	36	1	6
West Germanic	36	28	85	-	305	17	44	54	54	22	29	10
North Germanic	36	19	26	192	-	5	16	25	23	21	2	9
Baltic	29	32	23	26	24	-	90	40	46	19	45	6
Slavic	47	45	67	72	35	59	-	80	72	52	22	10
Indic	60	51	64	83	34	26	94	-	161	39	33	17
Iranian	89	41	86	61	43	21	69	224	-	45	25	44
Albanian	48	41	75	26	14	14	47	54	60	-	10	7
Greek	55	28	18	23	11	31	30	68	46	31	-	4
Armenian	43	7	42	22	12	10	20	74	77	21	6	-

Fig. 3 Total numbers of word borrowing events between each pair of language groups. For instance, 53 words of the Italic group were borrowed from the languages of the Celtic group; 10 highest values are highlighted in *dark grey* and 12 following highest values in light *grey*

	Celtic	Italic	French Iberian	W. Germanic	N. Germanic	Baltic	Slavic	Indic	Iranian	Albanian	Greek	Armenian
Celtic	-	3.98	4.36	5.4	4.26	2.65	2.63	4	5.27	5.57	3.07	9.84
Italic	3.6	-	18.99	1.78	1.76	1.66	1.74	2.2	2.03	3.35	1.84	0.27
French/Iberian	2.2	19.58	-	2.09	1.25	0.66	0.66	2.07	2.28	3.85	0.1	1.6
West Germanic	2.4	2.1	4.52	-	22.38	2.82	1.7	2.43	3.43	2.36	2.97	2.66
North Germanic	2.4	1.43	1.38	11.78	-	0.83	0.62	1.12	1.46	2.25	0.2	2.39
Baltic	1.93	2.4	1.22	1.6	1.76	-	3.48	1.8	2.92	2.03	4.61	1.6
Slavic	3.14	3.38	3.56	4.42	2.57	9.78	-	3.6	4.57	5.57	2.25	2.66
Indic	4	3.83	3.4	5.09	2.49	4.31	3.64	-	10.22	4.18	3.38	4.52
Iranian	5.94	3.08	4.57	3.74	3.15	3.48	2.67	10.07	-	4.82	2.56	11.7
Albanian	3.2	3.08	3.99	1.6	1.03	2.32	1.82	2.43	3.81	-	1.02	1.86
Greek	3.67	2.1	0.96	1.41	0.81	5.14	1.16	3.06	2.92	3.32	-	1.06
Armenian	2.87	0.53	2.23	1.35	0.88	1.66	0.77	3.33	4.89	2.25	0.61	-

Fig. 4 Percentages of words affected by borrowing from other groups. For instance, 3.98% of the words of the Italic group have the Celtic origin. The same color notations as in Fig. 3, were adopted here

cluster combining the Indic and Iranian groups has a sustained influence on the other groups. In the same way, we mapped the 10 most intensive transfers into the IE evolutionary tree (Fig. 5b). Some other high percentages (in light grey) can be explained either by well-known historical migration events (e.g., between the Armenian and Iranian groups) or should be investigated in more detail (e.g., between the Slavic and the Albanian groups). For instance, Armenian borrowed so many words from the Iranian languages that it was at first considered a part of the Indo-Iranian languages, and was not recognized as an independent group of the Indo-European languages for many decades (Waterman 1976) (see the value of 11.7% for the transfers form Iranian to Armenian in Fig. 4). On the other hand, Baltic languages are extremely well preserved, retaining archaic features similar to ancient Latin and Greek. Similarities of the Baltic languages to ancient Greek (see the value of 5.14% for Greek to Baltic in Fig. 4) and Sanskrit (see value of 4.31% for Indic to Baltic in

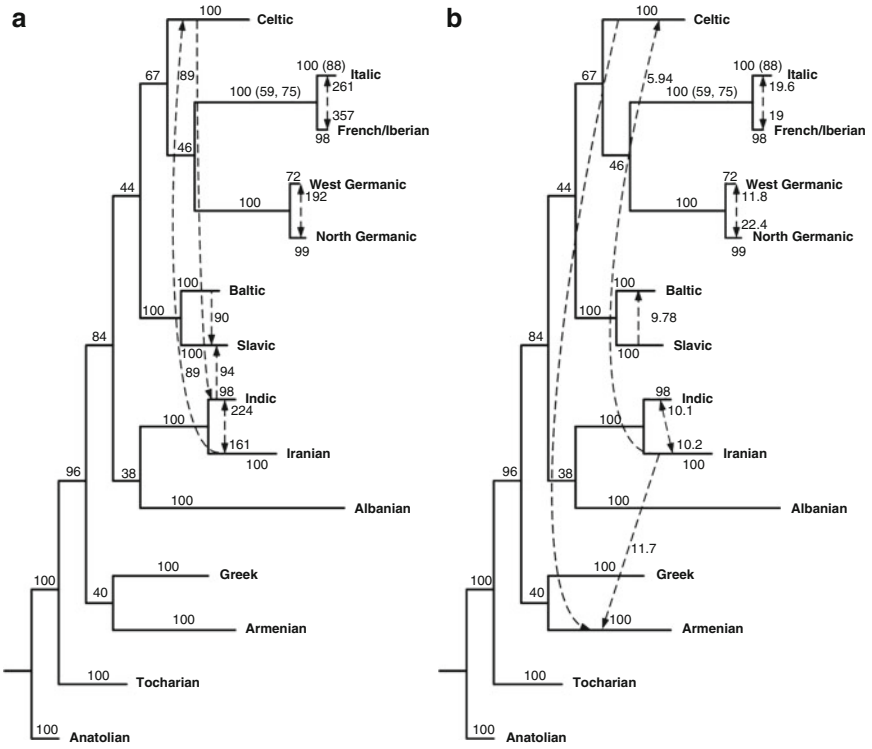


Fig. 5 Ten most frequent word exchanges between the IE language groups in terms of (**a**) total numbers of transferred words, and (**b**) percentages of affected words by group

Fig. 4) were noted long ago by Franz Bopp, the founder of comparative linguistic (Bopp 1867). Overally, 37% of the considered words were affected by borrowing from other language groups. The analogous results were obtained for the words of the lexical category (36.9%) and functional category (37.1%).

5 Conclusion

In this paper, we reconstructed a phylogenetic network of the Indo-European language family. The obtained network allowed us to represent the word borrowing events that have an important influence on the evolution of the IE languages. We found that 37% of the IE words have been affected by borrowing from other IE groups. Very similar results were obtained for the lexical and functional categories. This means that the word borrowing process does not depend on the broad lexical/functional category. However, the obtained result should be interpreted with caution because some of the word similarities, even for words belonging to different

language groups, can be due to accidental resemblance. In the future, we plan to conduct a refined study where the cognate content information (Dyen et al. 1997) will be taken into account. This should eliminate the impact of the accidental word similarities. We also found that the clusters combining the Indic and Iranian groups, and the Celtic, Italic, French/Iberian, West/North Germanic groups have much closer relationships than it is represented in the traditional IE tree (Gray and Atkinson 2003). This may be the evidence of a much closer common ancestry between these two clusters or of an intensive migration of the ancestors of the involved nations. In the future, it would be important to carry out a more comprehensive words borrowing analysis based on the 850 words of the Basic English (Ogden 1930). Basic English is an English-based controlled language created by Ogden (1930) (in essence, a simplified subset of English) as an international auxiliary language. Such a new analysis could help find more recent activities of borrowing. It would be also interesting to establish a parallel between each of the determined high word borrowing activities (see Figs. 3 and 4) and the historical events, external to the internal language systems, such as wars, migrations, or important commercial trades between related nations.

References

Atkinson, Q. D., & Gray, R. D. (2005). Curious parallels and curious connections: Phylogenetic thinking in biology and historical linguistics. *Systems biology, 54*, 513–526.

Barthelémy, J.-P., & Guénoche, A. (1991). *Trees and proximity representations*. New York: Wiley.

Bopp, F. (1867). *A Comparative Grammar of the Sanskrit, Zend, Greek, Latin, Lithuanian, Gothic, German, and Slavonic Languages*. (L. Eastwick, Trans). London: Madden and Malcolm, 1845–1856.

Diamond, J., & Bellwood, P. (2003). Farmers and their languages: The first expansions. *Science, 300*, 597–603.

Dyen, I., Kruskal, J. B., & Black, P. (1997). Comparative IE Database Collected by Isidore Dyen, from http://www.ntu.edu.au/education/langs/ielex/IE-RATE1. 1997.

Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature, 426*, 435–439.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*, 707–710.

Makarenkov, V., Boc, A., Delwiche, C. F., Diallo, A. B., & Philippe, H. (2006). New efficient algorithm for modeling partial and complete gene transfer scenarios. In V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Ziberna (Eds.), *IFCS 2006, Series: Studies in classification, data analysis, and knowledge organization, Springer Verlag*, 341–349.

Ogden, C. K. (1930). *Basic English: A general introduction with rules and grammar.* London: Paul Treber & Co., Ltd.

Pagel, M. (2000). *In time depth in historical linguistics.* In C. Renfrew, A. McMahon, and L. Trask, (Eds.) (pp. 189–207).

Rexová, K., Frynta, D., & Zrzavý, J. (2003). Cladistic analysis of languages: Indo-european classification based on lexicostatistical data. *Cladistics, 19*, 120–127.

Robinson, D. R. & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences, 53*, 131–147.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution, 4*, 406–425.

Swadesh, M. (1952). Lexico-statistical dating of prehistoric ethnic contacts: With special reference to north american indians and eskimos. In *Proceedings of the American Philosophical Society, 96*, 452–463.

Waterman, J. (1976). *A history of the German language.* Seattle, WA: University of Washington Press.

Parsing as Classification

Lidia Khmylko and Wolfgang Menzel

Abstract Dependency parsing can be cast as a classification problem over strings of observations. Compared to shallow processing tasks like tagging, parsing is a dynamic classification problem as no statically predefined set of classes exists and any class to be distinguished is composed of pairs from a given label set (syntactic function) and the available attachment points in the sentence, so that even the number of "classes" varies with the length of the input sentence. A number of fundamentally different approaches have been pursued to solve this classification task. They differ in the way they consider the context, in whether they apply machine learning approaches or not, and in the means they use to enforce the tree property of the resulting sentence structure. These differences eventually result in a different behavior on the same data making the paradigm an ideal testbed to apply different information fusion schemes for combined decision making.

1 Introduction

Syntactic parsing aims at determining structural properties among the word forms of a natural language utterance as a prerequisite for further processing, like information extraction, semantic interpretation, or machine translation. For this purpose, hierarchical descriptions are usually used.

Two different paradigms are mainly used: phrase or dependency structures. In phrase structure parsing, word forms are recursively grouped into increasingly larger constituents which carry a category taken from a finite inventory. In this sense phrase structure parsing can be seen as a classification problem, however, neither the number nor the local extensions of the objects to be classified (i.e. the constituents) is known beforehand.

L. Khmylko (✉)
Natural Language Systems Group, University of Hamburg, Vogt-Kölln-Straße 30,
D-22527 Hamburg, Germany
e-mail: khmylko@informatik.uni-hamburg.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_72,

Dependency structures renounce the phrasal nodes and express the syntactic relations in a sentence as edges of a directed graph the nodes of which are built of the word tokens. With an artificial root token, such a graph is acyclic and (weakly) connected, it also has to satisfy the single-head constraint and, thus, makes up a tree.

Dependency parsing much better fits the notion of classification because there the structural description can be broken down into a number of separate pieces of information which need to be assigned to each word form in the input utterance. Hence, it resembles the much simpler problem of tagging, for which a number of solution methods is known.

The goal of tagging is to uniquely assign a category (syntactic or semantic) $t(w_i)$ from a finite set T to each of the locally ambiguous word forms in the input utterance based on information about the individual context in which they occur. The influence of the context on the decision can either be captured by constraint-based (Karlsson et al. 1994), transformation-based (Brill 1995), or probabilistic multi-class classification (Brants 2000) methods.

To extend the idea of tagging to dependency parsing the representation needs to be changed in a way that it allows to model linguistic dependency as the attachment of one word form to another. That requires replacing the fixed predefined set of categories by a set of positional indices which can be interpreted as attaching the given word form to another one in the input utterance $d(w_i) = j, i \neq j$, $i = 1, \ldots, N$, $j = 0, \ldots, N$, where a special position (in our case the one with index 0 is used to provide an attachment point for the top-most node of the dependency tree. Sometimes, the problem is further restricted to projective dependency trees which have to obey the additional constraint on positional indices: $i < j < k \wedge (d(w_i) = k \vee d(w_k) = i \wedge d(w_j) = l \rightarrow i \leq l \leq k))$, i.e. a dependency edge is not allowed to cross the projection line of another word form. Such a representation, also called bare dependencies, is not sufficient for many applications. Usually, one wants to also assign a (syntactic) function, like subject or genitive modifier, to the structural relationships, i.e. a label, again taken from a finite inventory $d(w_i) = (j, l), i \neq j, i = 1, \ldots, N, j = 0, \ldots, N, l \in L$.

These two assignment tasks are taken as standard scenarios for evaluating a dependency parser with respect to its output quality. Yet another refinement could be introduced, since a word form typically has more than a single lexical reading (morpho-syntactic, syntactic, or semantic). The parsing problem can again be extended to also regard which of the available readings from a word-form specific set R_i was considered when choosing the attachment point and the label $d(w_i) = (j, l, r), i \neq j, i = 1, \ldots, N, j = 0, \ldots, N, l \in L, r \in R_i$.

Although the final description has grown in complexity, the fundamental similarity to a classification problem is preserved: The set of possible (complex) categories and the objects to be classified are known in advance. Therefore, a similar variety of solution methods as for the tagging problem is available. Also, tagging and parsing share the same underlying difficulty that a reliable decision cannot be taken locally but has to consider the context, both in terms of input word forms and their structural relationships. Even worse it turned out that in most cases the complete utterance will

be needed instead of a limited window of word forms mostly sufficient for tagging as additional global well-formedness conditions have to be observed. Thus, in case of dependency parsing the problem grows to be multi-class structured classification problem.

Recently, three systems for dependency parsing have been particularly successful: WCDG (Foth and Menzel 2006), MSTParser (McDonald et al. 2006) and MaltParser (Nivre et al. 2007). The following sections primarily discuss these three. Subsequently possible ways of parser combination are analyzed.

2 WCDG

Weighted Constraint Dependency Grammar (WCDG) rests on the formalism of a Constraint Dependency Grammar extended with weights by Schröder (2002) to better deal with structural ambiguities. Well-formed structures are described in this grammar entirely by constraints. The relative importance of constraints is expressed by a weight lying in the interval between zero and one, a lower value makes a constraint more prohibitive. The parse found by the system is analyzed for constraint violations between the structure and the rules expressed by constraints. The score of an analysis is the product of all the weights for constraint violations occurring in the parse.

Current WCDG for German comprises about 1,000 constraints and can be freely obtained from http://nats-www.informatik.uni-hamburg.de/view/CDG/DownloadPage.

WCDG treats the parsing problem as a Constraint Satisfaction Problem and searches for an analysis with the highest score. Unfortunately, applying a complete search is intractable for this problem type, but efficient heuristics exist. The most reliable method has proven to be the transformation-based solution. Starting with an initial guess about the optimal tree, changes of labels, subordinations, or lexical variants are applied, whereby the constraint violations are used as a control mechanism guiding the transformation process (Foth et al. 2000). If the rearrangement of the tree structure has not been successful, increasingly long transformation sequences are tried out.

Resolving the optimum is not guaranteed by the transformation-based search, but this method does provide additional benefits. It is not only more resource efficient than the complete search, but it can be interrupted at any time and will always return an analysis with a list of constraint violations that could not be disposed of. The algorithm terminates on its own if the parse does not violate any constraints above a predefined threshold or if an optional timeout is reached.

Although the approach employed by WCDG has a number of limitations as such, it can serve as a framework for integrating contributions from external predictor components in a soft manner.

Five additional statistical components have been previously added to WCDG – tagger, chunker, supertagger, PP attacher, shift-reduce oracle – even though their

accuracy lies mostly, with the exception of the tagger, below that of the parser itself, WCDG not only avoids error propagation successfully, but its performance improves slightly with each component added so that structural accuracy grows from 89.7% for the combination with the tagger alone to 92.5% for the experiment in which all five predictors interact (Foth and Menzel 2006). Labeled accuracy increases thereby from 87.9% to 91.1%.

3 MSTParser

MSTParser (McDonald et al. 2006) is a state-of-the-art language independent data-driven parser. It is freely available from http://sourceforge.net/projects/mstparser. MSTParser successfully combines discriminative methods for structured classification with graph-based solution methods for the parsing problem.

In this edge-factored graph-based model, each edge of the dependency graph is assigned a real-valued score that expresses the likelihood of creating a dependency edge between two words. The score of the graph is defined as the sum of its edge scores. The parsing problem becomes equivalent to finding the highest scoring directed spanning tree in the complete graph over the given sentence, and the correct parse can be obtained by searching the space of valid dependency graphs for a tree with a maximum score. The scoring function for edges is obtained by the Margin Infused Relaxed Algorithm (MIRA), an online maximum margin learning similar to Perceptron, which is extended for structured multi-class classification (Taskar 2004).

Efficient parsing algorithms have been found for both projective (Eisner 1996) and non-projective (Chu and Liu 1965; Edmonds 1967) dependency trees. When only features over single edges are taken into account, the complexity of the non-projective parsing falls to unprecedented $O(n^2)$.

The above mentioned parsing algorithms only deduces the bare dependency structure, labeling is applied to it in the second stage. Such an exclusion deprives the attachment stage of important label cues.

While the inclusion of more than one edge into the scoring function is desired to extend the locality of the context on which the decision is made, already considering adjacent edges in addition to single edges makes the non-projective parsing intractable. Thus, an approximate, but efficient algorithm based on exhaustive search with tree-to-tree transformation similar to the WCDG transformation-based solution is provided for this case (McDonald et al. 2006).

The average structural and labeled accuracy that MSTParser shows over thirteen languages (McDonald et al. 2006) are 87.0% and 80.8% respectively. Only for Arabic, Turkish and Slovene, the results fall below 80%, and for around half of the evaluated languages it achieves structural accuracy over 90%.

The parsing model of MSTParser has the advantage that it can be trained globally and eventually be applied with an exact inference algorithm. On the other hand, the parser has only limited access to the history of parsing decisions. To avoid

complexity problems, the scores (and the feature representations) are restricted to a single edge or adjacent edges.

4 MaltParser

MaltParser (Nivre et al. 2007) is another language-independent parsing system for data-driven dependency parsing (freely available for research and education from http://w3.msi.vxu.se/users/jha/maltparser/index.html).

MaltParser is a shift-reduce type parser which works in a single pass from left to right through the sentence maintaining two main data structures – a queue of remaining input tokens and a stack storing partially processed tokens. At every processing step, it deterministically decides about the next elementary parser action: shift to the stack, reduce or attach the top word in the stack to the left or right.

The words from the queue are successively shifted onto the stack until the top word can be attached to the next word from the queue as its left or right dependent. Reduce action pops the stack when the top word has been attached on the left to a node other than the root and has already found all its right dependents. Shift action is aimed at the nodes that have their heads on the right or remain attached to the root node ever since initialization.

The shift-reduce algorithm is restricted to finding projective dependency structures and attaching to the right is performed together with popping the stack, because all left and right dependents must have been already identified due to projectivity. Attaching to the left immediately pushes the next word from the queue onto the stack since no new left dependents are to be expected at this point, but new dependents on the right may still exist and so the word cannot be popped.

In general, more than one action may be applied at each processing step, thus, to make a deterministic choice the parser relies on oracle predictions. It uses memory-based learning (MBL) employing history-based feature models and discriminative machine learning. The main idea of MBL is that parsing actions that have to be applied in some parsing state are similar to other that were applied previously in similar parser configurations and thus the memorized solutions may be reused.

While MSTParser considers the global context when scoring the different hypotheses, MaltParser conditions its decisions on its parsing history, i.e. the parsing actions chosen so far, it approximates a globally optimal solution by applying a series of locally optimal decisions.

To deriving an analysis, MaltParser needs time linear in the length of the sentence which is very efficient. The tree property of the resulting structure is guaranteed by the parsing algorithm implicitly. Moreover, labeling is integrated into the attachment actions.

MaltParser shows unlabeled dependency accuracy above 80% (Nivre et al. 2007) for 10 languages, whereby it shows the best result of 88.1% for English and German. Its labeled accuracy ranges from 69.0% for Turkish to 86.3% for English with the majority of languages lying above 75%.

Dealing with non-projective trees comes at the expense of the growing complexity that becomes quadratic in the length of the sentence if the incremental algorithm by Covington (2001) is used. The stack is then replaced by an open list in which each word can be attached to the next word from the queue and thus non-projective structures may be derived.

Another technique to relax the projectivity restriction is the pseudo-projective parsing proposed by Nivre and Nilsson (2005). It extends the main algorithm with two additional graph transformation stages. First, the parser input is pre-processed so that non-projective structures are transformed into projective ones by applying minimal possible changes to the tree. During an additional post-processing phase, non-projective structures are being recovered by performing exhaustive search for the real head constrained by an extended arc label.

5 Parser Combination

Traditionally, approaches to parser combination haven been mainly based on the idea of a post-hoc selection which can be carried out for either complete parses, or individual constituents and dependency edges, respectively (Henderson and Brill 1999). The selection component itself is based on heuristic procedures, like a majority vote. Alternatively, a second-level classifier is trained to decide which component to trust under which conditions, an approach often referred to as classifier stacking (Zeman and Žabokrtský 2005). Since global consistency criteria are crucial for deciding about the final sentence structure, a large amount of information about the context of a partial structure needs to be considered. Therefore, approaches seem to be more promising which instead of relying on an ad hoc selection of features use all the available information and integrate it by means of a single global decision criterion. This can be achieved at three different points in time: prior, during, or past parsing.

Pre-parsing integration is taking place at training time. The training data for one classifier is re-annotated with features taken from the parsing result of the other one and vice versa. Nivre and McDonald (2008) combined MaltParser and MSTParser this way. They trained a new model, which, in addition to the features from the input sentence, uses features provided by the predictor model. The decision on the optimal parse is still taken by the original parsing algorithm, without resorting to any local selection heuristics. Since the two parsers are conditioned in quite different ways (parse history vs. parse context) they exhibit a remarkable complementary behavior (McDonald and Nivre 2007). Accordingly, significant mutual benefits have been observed. Averaged over data from 13 different languages they report an improvement of labelled accuracy from 80.83% to 82.53% using MSTparser enriched with MaltParser results, and from 80.74% to 82.01% in the other direction. Note however, that one of the major benefits of MaltParser, its incremental left-to-right processing, is sacrificed under such a combination scheme.

For systems based on manually developed grammars like WCDG, a pre-parsing combination is not possible. Its equivalent consists in making the output of the second parser immediately available to the decision procedure as additional cues. WCDG's predictor mechanism is perfectly suited for that purpose. Additional constraints check the compatibility of the current hypothesis with the external prediction and impose a penalty if necessary. The advantage of such a combination consists in the high degree of independence from the additional information: Even if no external predictions are available, the original systems remains fully functional. So far, however, no online integration of external predictions is possible. MSTParser has comparably high performance to WCDG. On the German test data MSTParser shows 90.5% structural and 87.5% labeled accuracy. Still, with MSTParser as an oracle, WCDG achieves higher performance than each of the combined components in isolation: 92.9% and 91.3% structural and labeled accuracy respectively.

In two post-parsing experiments, Sagae and Lavie (2006) combined a number of dependency and constituent parsers, respectively. They created a new weighted search space from the results of the individual component parsers using different weighting schemes for the candidates. They then re-parsed this search space and were able to improve the unlabeled accuracy of dependency structures to 92.7% using an ensemble of four parsers of English with accuracies between 98.6% and 91.0%. For phrase structure trees they combined five parsers with F-scores between 86.7% and 91.0% and reached an overall F-score of 92.1%.

6 Conclusion

Dependency structures represent a major benefit to treat parsing as a classification problem. Dependency structures can be specified in a more narrow locality in comparison to phrase structures as per word relationships can be modeled a straightforward way. Besides, lexical information is integrated into the structure.

Dependency parsing has a certain similarity to tagging, both of which can be considered a dynamic classification problem. It turns out, however, that local criteria for deciding on the best attachment point and label are not sufficient either for parsing itself or for parser integration as the tree constraint should be additionally met. Still different structured classification schemes are available.

References

Brants, T. (2000). TnT – A statistical part-of-speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing* (pp. 224–231).

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 543–565.

Chu, Y. J., & Liu, T. H. (1965). On the shortest arborescence of a directed graph. In *Science Sinica, 14*, 1396–1400.

Covington, M. A. (2001). A fundamental algorithm for dependency parsing. In *Proceedings of the 39th Annual ACM Southeast Conference* (pp. 95–102).

Edmonds, J. (1967). Optimum branchings. In *Journal of Research of the National Bureau of Standards, 71B*, 233–240.

Eisner, J. (1996). Three new probabilistic models for dependency parsing: An exploration. In *Proceedings of the International Conference on Computational Linguistics* (pp. 340–345).

Foth, K. A., Menzel, W., & Schröder, I. (2000). A transformation-based parsing technique with anytime properties. In *4th International Workshop on Parsing Technologies, IWPT-2000* (pp. 89–100).

Foth, K. A., & Menzel, W. (2006). Hybrid parsing: Using probabilistic models as predictors for a symbolic parser. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics* (pp. 321–328).

Henderson, J. C., & Brill, E. (1999). Exploiting diversity in natural language processing: Combining parsers. In *Proceedings of the 4th Conference on EMNLP* (pp. 187–194).

Karlsson, F., Voutilainen, A., Heikkila, J., & Anttila, A. (1994). *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Berlin, New York: Mouton De Gruyter.

McDonald, R., Lerman, K., & Pereira, F. (2006). Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of CoNLL-X 2006* (pp. 216–220).

McDonald, R., & Nivre, J. (2007). Characterizing the errors of data-driven dependency parsing models. In *Proceedings of EMNLP-CoNLL 2007* (pp. 122–131).

Nivre, J., Hall, J., Nilsson, J., Eryiğit, G., & Marinov, S. (2007). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering, 13*(2), 95–135.

Nivre, J., & McDonald, R. (2008). Integrating graph-based and transition-based dependency parsers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics with the Human Language Technology Conference* (pp. 950–958).

Nivre, J., & Nilsson, J. (2005). Pseudo-projective dependency parsing. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (pp. 99–106).

Sagae, K., & Lavie, A. (2006). Parser combinations by reparsing. In *Proceedings of HLT-NAACL* (pp. 129–132).

Schröder, I. (2002). *Natural language parsing with graded constraints*. PhD thesis, Department of Computer Science, University of Hamburg, Germany.

Taskar, B. (2004). *Learning structured prediction models: A large margin approach*, PhD Dissertation, Stanford University.

Zeman, D., & Žabokrtský, Z. (2005). Improving parsing accuracy by combining diverse dependency parsers. In *Proceedings of the 9th International Workshop on Parsing Technologies* (pp. 171–178), Vancouver, BC, Canada.

Comparing the Stability of Clustering Results of Dialect Data Based on Several Distance Matrices

Edgar Haimerl and Hans-Joachim Mucha

Abstract How can the investigation of the hidden structure in a proximity matrix be condensed and targeted more towards domain expert knowledge? A data matrix that combines different statistical measures of the proximity matrix under investigation is proposed and evaluated. In order to validate the outcome and to measure how well this requirement is met the original and the compound matrix have to be compared. This is done by applying algorithms to determine cluster stabilities as introduced in Mucha and Haimerl (2005): A simulation algorithm finds the best number of clusters, calculates the stability of clusters found in hierarchical cluster analysis, and at the most detailed level calculates the rate of recovery by which an element can be reassigned to the same cluster in successive classifications of bootstrap samples. Both the cluster stability and the consistency of the clustering results with expert expectations prove the advantage of the compound matrix over the generally used proximity matrix.

1 Introduction

First we are going to explain the background for establishing a "compound" distance matrix based on some statistical parameters of an original proximity matrix. Here proximities are the general term for pairwise distances or similarities. It is easy to switch between distances d and similarities s by d = max(s)-s (or in the case of our data sample by d = 100-s). The main part of the paper consists of comparing these two matrices via hierarchical cluster analysis based on cluster validation by subsampling as introduced in Mucha and Haimerl (2005). Here, the simulation algorithms determine the number of clusters, the stability of clusters found in hierarchical cluster analysis, and at the most detailed level the rate of recovery by which an element can be reassigned to the same cluster in successive classifications of

E. Haimerl (✉)
Institut für Romanistik, Universität Salzburg, Akademiestraße 24, A-5020 Salzburg, Austria
e-mail: Edgar@Haimerl.eu

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_73,

bootstrap samples. To proof this methodological concept it is applied to quantitative linguistics data. In order to ease the comparison, the same data set as in Mucha and Haimerl (2005), Haimerl and Mucha (2007) will be used – data from dialect research of Northern Italy (Goebl 1998).

2 The Compound Matrix

Each row in a similarity matrix contains the similarity of one item (location) under investigation to all other items in the corpus. Applying techniques of descriptive statistics to each row yields characteristic values which differentiate the array of values in one row from other arrays and thus indirectly one item from all other items. Figures 1–3 show the nonparametric density estimation of the array of pairwise similarities for locations A8, A77, and A122. They look quite different.

The values of the measures of range, central tendency and variability of a row in the similarity matrix can be considered as a vector characterising the row (see Table 1). A data matrix consisting of these vectors of statistical measures of the original proximity matrix defines a "higher level" data matrix representing the hidden structure on a more abstract level. This compound data matrix serves as input in further analysis, e.g. an Euclidean distance matrix can be calculated and used as input for further cluster analysis.

The compound matrix is a more abstract view of the structure in the data under investigation than the initial proximity matrix with values which have been

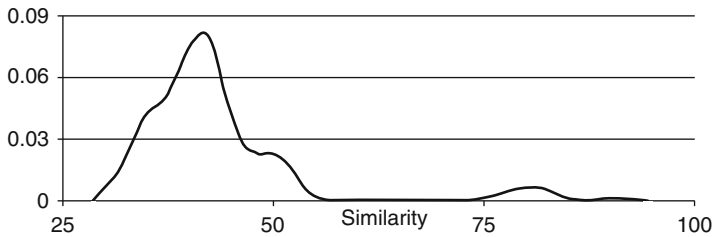


Fig. 1 Nonparametric density estimation of pairwise similarities of location A8

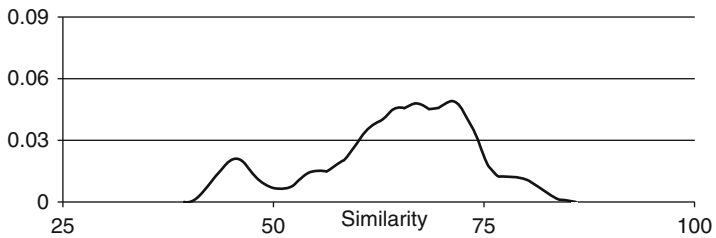


Fig. 2 Nonparametric density estimation of pairwise similarities of location A77

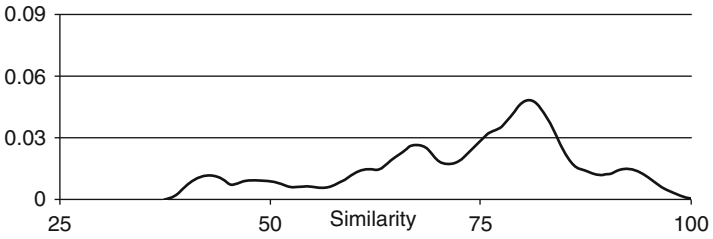


Fig. 3 Nonparametric density estimation of pairwise similarities of location A122

Table 1 Statistical measures of locations (excerpt)

Location	Minimum	Mean	Median	Maximum	Std.Dev.	Skewness
A1	35.51	46.81	44.97	87.36	9.60	2.64
...	...	...	...	...	...	...
A8	31.80	43.14	41.19	91.20	10.09	2.68
...	...	...	...	...	...	...
A77	43.01	64.21	66.01	82.55	9.45	−0.61
...	...	...	...	...	...	...
A122	41.00	72.92	76.13	96.83	13.64	−0.66
...	...	...	...	...	...	...

calculated e.g. with the simple matching coefficient. This abstraction has to be investigated in more detail in the context of the linguistic data. Similarity matrices of spacial data can be visualized in VDM (Visual DialectoMetry) as reference maps. For reference maps of the ALD area, explanations of the maps and interpretations see (Bauer 2003, 2009; Haimerl n.d.). As a first approach reference maps based on the RIV (relative identity value see Goebl 1984, p. 74) similarity matrix and reference maps based on the compound matrix can be visualized side by side in two VDM windows. Figure 4 shows the reference map of Brail (location A 8) of the compound matrix. Locations with high similarity are dark. What strikes the eye at first sight is that all locations in Grischun and in the three norther Dolomitic Ladin valleys and some locations in Friuli – presented as dark and medium gray polygones – share very high similarity though they are spatially far apart. But from a linguistic view this makes sense as these are the Raeto-Romance language areas. Further differences are the transition zones (light gray and hatched double diagonally) where the grouping is scattered and no clear structures are observable.

This first visual impressions are depend by the map showing the Pearson's product-moment coefficient for each location (Fig. 5). We see low correlations (horizontally hatched) especially at the border areas between the Raeto-Romance areas in the North and the Italien language areas in the South whereas in Veneto the correlation is very high (dark gray).

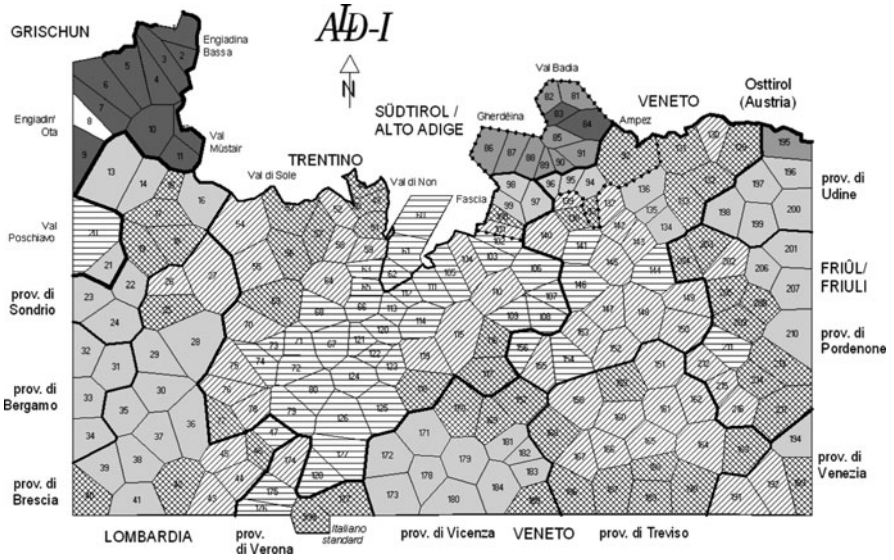


Fig. 4 Reference map of location A8 (Brail)

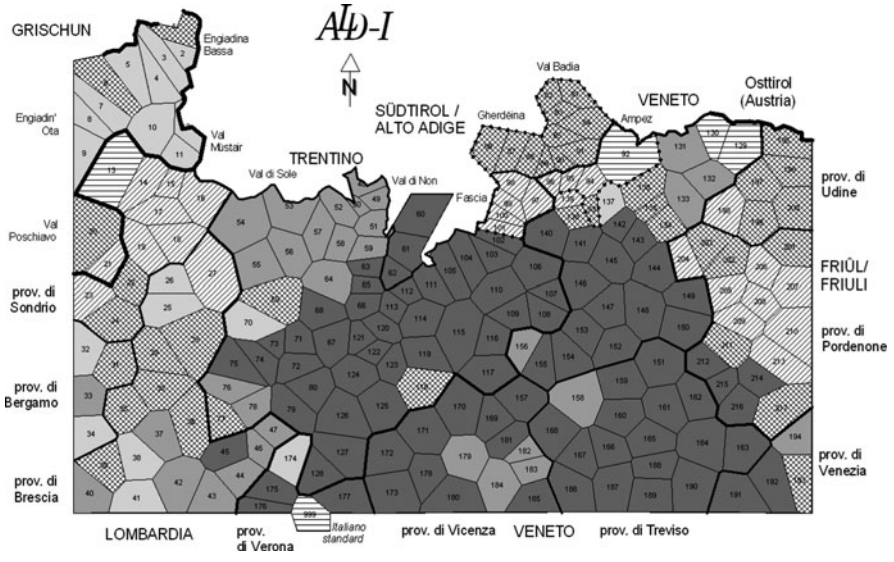


Fig. 5 Mapping of the product-moment coefficient for each location

3 Why Use these Statistical Measures for Linguistic Data?

The statistical measures of rows in the proximity matrix describe these items relative to the other items (rows) in statistical terms. In addition to this statistical description dialectometric experts did come up with content and topic oriented explanations. In order to condense the structure hidden in the RIV similarity matrix and to target the information toward the knowledge and expectations of linguists and dialectologists the most frequently discussed statistical measures are applied:

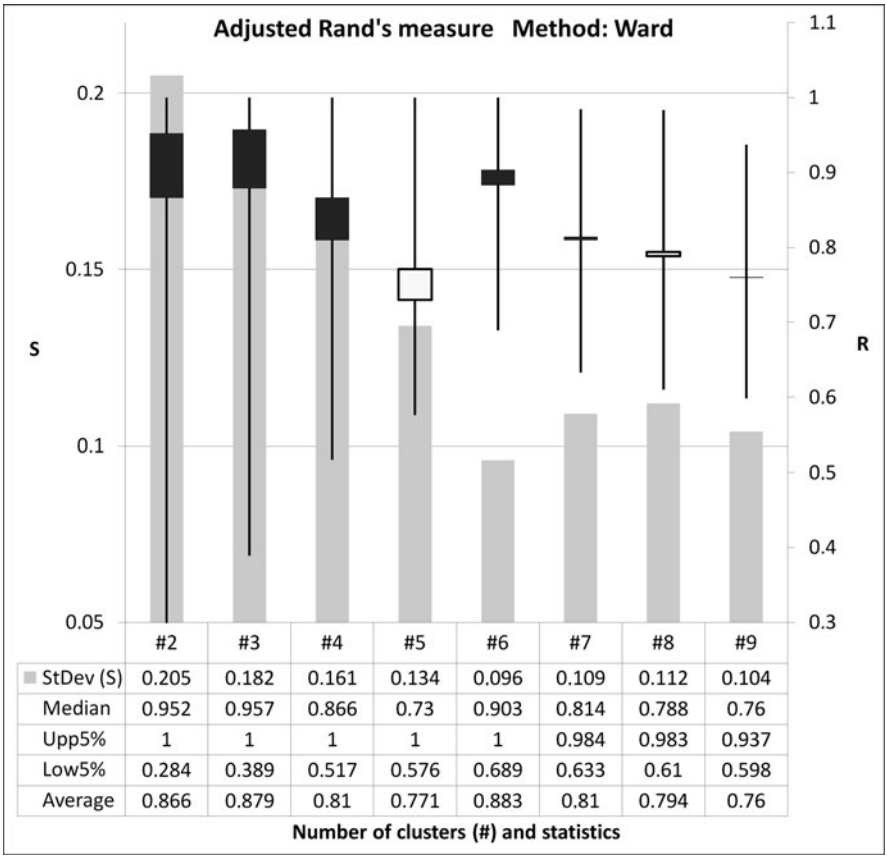
- Minimum: You can think of the location with the smallest similarity to all other locations as the most disliked location - it is the chief opponent. (see Goebl 1998; Bauer 2004).
- Maximum: The location with the highest similarity to all other locations has a lot of friends with which it is closely related. This is characteristic for locations in dialect kernels. (see Goebl 2001, p. 183).
- Mean and median: These measures of central tendency express how good a location is cross-linked with the environment under investigation. High mean and median values indicate that this location has only few locations with low similarity; thus this location is tightly cross-linked with its neighbors. (see Goebl 1984, p. 148).
- Standard deviation: this measure of dispersion sheds light on the spacial dispersion of language systems. (see Goebl 2004, p. 266).
- Skewness is of special interest for the linguistic interpretation as this measure of distribution gives information about linguistic comprise or exchange – "Sprachausgleich". (see Goebl 2006, p. 241).

Each of these values fills one column in the compound matrix; the values are normalized and disturbing weights from high correlations between statistical measures are eliminated. Certainly many more measures like higher order statistics could be applied, but we want to restrict the investigation to those measures for which linguistic interpretations are published.

4 Comparing Hierarchical Cluster Results

4.1 Cluster Stability Results

The comparison of these two proximity matrices via hierarchical cluster analysis is based on cluster validation by subsampling as introduced in Mucha and Haimerl (2005). Here, the simulation algorithms determine the number of clusters, the stability of clusters found in hierarchical cluster analysis, and at the most detailed level the rate of recovery by which an element can be reassigned to the same cluster in successive classifications of bootstrap samples. As already seen in Haimerl and Mucha



	#2	#3	#4	#5	#6	#7	#8	#9
StDev (S)	0.205	0.182	0.161	0.134	0.096	0.109	0.112	0.104
Median	0.952	0.957	0.866	0.73	0.903	0.814	0.788	0.76
Upp5%	1	1	1	1	1	0.984	0.983	0.937
Low5%	0.284	0.389	0.517	0.576	0.689	0.633	0.61	0.598
Average	0.866	0.879	0.81	0.771	0.883	0.81	0.794	0.76

Fig. 6 Output of the validation tool

(2007), the adjusted Rand measure yields reliable results for cluster comparison. Figure 6 shows the result of the validation.

Comparing the adjusted Rand values of HCA based on the compound matrix with the values of HCA of the RIV matrix (see Mucha and Haimerl 2005, Fig. 2) yields that

- The adjusted Rand values of the compound matrix are higher – max value 0.9 versus 0.8 for the RIV matrix.
- There is a clear vote for six partitions, whereas the RIV Rand values vote for seven partitions but with less good distinction.

4.2 Interpretation of the Dialect clusters

The stability of the HCA result based on the compound matrix does not guarantee that the clusters are meaningful and valuable for domain experts. The clustering

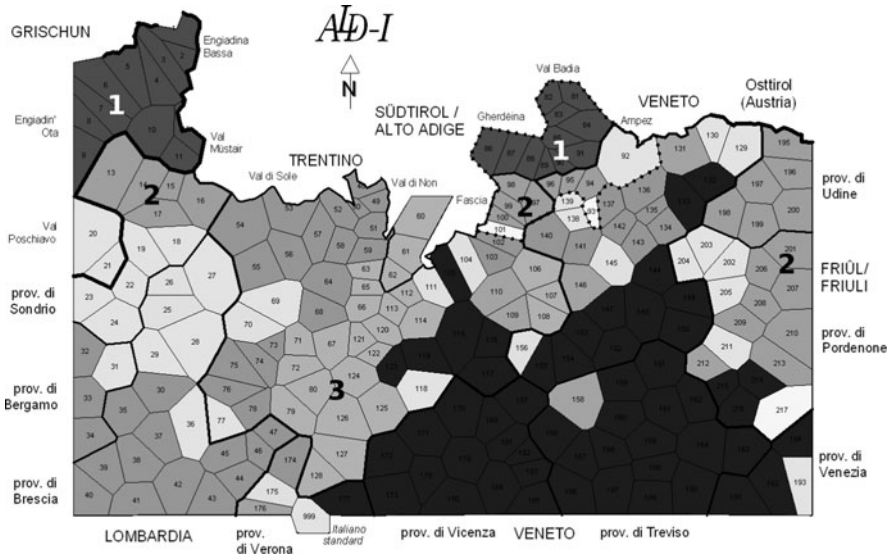


Fig. 7 HCA results based on the compound matrix

results can be mapped as Voronoi map (as shown in Fig. 7); one polygon for each location in one color for each cluster. Does this mapping of the clusters make sense to dialect experts of Northern Italy? Three out of the six clusters are very stable – the recovery rate is better than 95% for all of the clustered locations. Cluster (1) connects quite distant regions: Grischun/Switzerland with the three norther Dolomitic Ladin valleys. Cluster (2) extends over three dispersed regions: direct neighbors to Grischun, the two southern Ladin valleys and Friul. These are in fact very interesting dialect clusters which match closely with the Raeto-Romance concept which goes back to Ascolli. Cluster (3) is not surprising and has been found in many prior clusterings of the area under investigation.

What is unique for the HCA result based on the compound matrix is that locally dispersed language structures are joined into one cluster. This becomes evident when Fig. 7 is compared with the result from the same HCA algorithm applied to the RIV matrix directly as discussed in Mucha and Haimerl (2005). The fact that the cluster partition based on the compound matrix is less biased by spacial distances can be explained: The compound matrix uses the characteristics of the locations as derived from their statistical measures. For example the locations which share all characteristics "chief opponent – not well cross linked – resists to language compromise" end up in one group. There is no indirect spacial distance in the compound data matrix and derived proximity matrices.

References

Bauer, R. (2003). Sguardo dialettometrico su alcune zone di transizione dell'Italia nord-orientale (lombardo vs. trentino vs. veneto). In R. Bombi & F. Fusco (Eds.), *Parallela X. Sguardi reciproci. Vicende linguistiche e culturali dell'area italofona e germanofona. Atti del Decimo Incontro italo-austriaco dei linguisti* (pp. 93–119). Udine: Forum Editrice.

Bauer, R. (2004). Dialekte – Dialektmerkmale – dialektale Spannungen. Von 'Cliquen', 'Störenfrieden' und 'Sündenböcken' im Netz des dolomitenladinischen Sprachatlasses ALD-I. In *Ladinia, XXVIII*, 201–242.

Bauer, R. (2009). FWF-Forschungsprojekt P14566-G01; ALD-DM. From http://ald.sbg.ac.at/ald/alddm/

Goebl, H. *Dialektometrie; Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie.* Wien: Verlag der Öst. Akademie der Wissenschaften.

Goebl, H. (1984). *Dialektometrische Studien anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF.* Bd.1-3, Tübingen.

Goebl, H. (Ed.). (1998). *Atlante linguistico del ladino dolomitico e dei dialetti limitrofi I (ALD I) – Sprachatlas des Dolomitenladinischen und angrenzender Dialekte I* (Vol. 7). Wiesbaden: Dr. Ludwig Reichert Verlag.

Goebl, H. (1998). On the nature of tension in dialectal networks; a proposal for interdisciplinary discussion, In: *Systems; New Pradigms for the Human Sciences,* 550–571.

Goebl, H. (2001). Skripta, Schreiblandschaften und Standardisierungstendenzen. In K. Gäertner e.a. (Eds.), *Urkundensprachen im Grenzbereich von Germania und Romania im 13. und 14. Jh.*, Trier, 169–221.

Goebl, H. (2004). Sprache, Sprecher und Raum: Eine kurze Darstellung der Dialektometrie. In *Mitteilung der Österreichischen Geographischen Gesellschaft,* 247–286.

Goebl, H. (2006). Metasprachliche Kon- und Divergenzen im Bereich der Sprachlandschaft Ladinia. In *Ladinia*, 223–283.

Haimerl, E. (n.d.). *Das Dialektometrieprojekt der Universität Salzburg. (in German and English).* http://ald.sbg.ac.at/dm

Haimerl, E., & Mucha, H.-J. (2007). Comparing the stability of different clustering results of dialect data. In R. Decker & H.-J. Lenz (Eds.), *Advances in data analysis* (pp. 619–626). Berlin: Springer.

Mucha, H.-J., & Haimerl, E. (2005). Automatic validation of hierarchical cluster analysis with application in dialectometry. In C. Weihs & W. Gaul (Eds.), *Classification – the ubiquitous challenge* (pp. 513–520). Berlin: Springer.

Marketing and Regional Sales: Evaluation of Expenditure Strategies by Spatial Sales Response Functions

Daniel Baier and Wolfgang Polasek

Abstract Non-linear production functions are a common basis for modelling regional sales responses to marketing expenditures. A recent article of Kao et al. (Evaluating the effectiveness of marketing expenditures. Working Paper, 2005) suggests to use such models to estimate the effectiveness of marketing strategies. In this paper the underlying approach is extended: Firstly, a spatial component is explicitly modelled in the production function, and secondly, a hierarchical approach in the clustering of regional sales is used. The developed Cross Sectional Sales Response (CSSR) models use Stochastic Partial Derivatives (SPD) constraints. They are tested using synthetic and pharma marketing data.

1 Introduction

The sales and cost effectiveness of marketing instruments has gained considerable attention in recent years. So, e.g., in the pharmaceutical industry, more and more companies are troubled by the effectiveness of their regional salespersons who personally visit physicians for explaining and promoting the company's products. Instead, they consider alternative instruments like, e.g., phone calls, one-to-one-video-conferencing, interactive websites, or online community events (see, e.g., Lerer 2002; Queitsch and Baier 2005).

However, in order to estimate the sales effectiveness of marketing instruments and strategies, a proper sales response modelling of these marketing instruments is needed. Here, recently, Kao et al. (2005) have proposed to use a class of non-linear production functions under optimization constraints for this purpose: Using x and z for (time-dependent) marketing expenditures and y for (time-dependent) sales their basic modelling assumption is the traditional multiplicative model

$$y = \gamma x^{\beta_1} z^{\beta_2} e^{\epsilon}. \tag{1}$$

D. Baier (✉)
Institute of Business Adminstration and Economics, BTU Cottbus, Postbox 101344, D-03013 Cottbus
e-mail: daniel.baier@tu-cottbus.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization, DOI 10.1007/978-3-642-10745-0_74,

By assuming that the company distributes its expenditures in the theoretically optimal sense that marginal returns are equal across instruments ($\partial y/\partial x = \partial y/\partial z$) Kao et al. estimate valid model parameters from observed (y, x, z)-values.

In this approach we extend this approach in two directions: First we model explicitly a spatial component in the production function and secondly, we explore the use of a hierarchical model in the clustering in order to optimize the geographic cost-effectiveness ratio of marketing strategies. We propose a new class of Cross Sectional Sales Response (CSSR) models. The goal is to test the effectiveness of existing regional marketing expenditures and to suggest new expenditure patterns. The hierarchical extension of the model is in the spirit of Rossi et al. (2005) and follows the idea, that the sales elasticities can vary geographically across macro-regions. While the original model of Kao et al. (2005) is a panel model that estimates the response parameters across time, the developed CSSR model can be estimated only on data from few time points. In Sect. 2 the basic CSSR model and the MCMC estimation procedure is developed and in Sect. 3 a spatial dimension is added. The last section concludes.

2 Cross Sectional Sales Response Models

2.1 The Basic CSSR Model

Starting from $y = \gamma x^{\beta_1} z^{\beta_2} e^{\epsilon}$, taking logs, using $\beta = (\beta_0, \beta_1, \beta_2)'$ with $\beta_0 = \log(\gamma)$, the CSSR model (in the following shortly: SRF for sales response function) with partial derivative restrictions is defined as

$$ln\ y \sim N(X\beta, \sigma_y^2 I_n). \tag{2}$$

This homoscedastic log-linear model has the conditional mean $\mu_y = X\beta$. Adding the partial derivative restrictions for the two regressors, which imposes the theoretical optimality conditions that the marginal allocations should be equal across units, in a stochastic way we obtain

$$\begin{aligned} ln\ x &\sim N(\mu_x, \sigma_x^2 I_n), \\ ln\ z &\sim N(\mu_z, \sigma_z^2 I_n) \end{aligned}$$

where the variances control the tightness of the optimality constraints: larger variances allow for more deviations from the optimal strategy. The conditional means $\mu_x = \mu_x(\beta, \lambda)$ and $\mu_z = \mu_z(\beta, \lambda)$ are given by

$$\begin{aligned} \mu_x &= (\beta_0 + ln\ \beta_1 - \lambda_1 + \beta_2 ln\ z)/(1 - \beta_1), \\ \mu_z &= (\beta_0 + ln\ \beta_2 - \lambda_2 + \beta_1 ln\ x)/(1 - \beta_2). \end{aligned}$$

This follows from both partial derivatives:

$$\partial y/\partial x = y_x = \beta_0\beta_1 x^{\beta_1 - 1} z^{\beta_2}$$
$$\partial y/\partial z = y_z = \beta_0\beta_2 x^{\beta_1} z^{\beta_2 - 1}.$$

Since x and z are fully observed quantities (like money expenses or sales efforts via local and global advertising), these restrictions take specific but known values for each observation, if the parameters of the model (β, σ_y^2) are fully known. Now we assume that the model can be estimated by imposing stochastic partial derivatives (SPD) constraints in the following form:

$$log(y_x) \sim N[\lambda_1, \tau_1^2]$$
$$log(y_z) \sim N[\lambda_2, \tau_2^2]$$

The λ_i's could be interpreted as some kind of average utility level of the sales responses while the τ_i^2's take the role of tightness parameters across observations in the sample. It seems reasonable to fix them as known hyper-parameters and to estimate the average marginal utilities λ_i's.

A further aspect of the SPD constraints are that by including marginal utility demons to a SRF we actually endogenize the inputs of the SRF and more complicated estimation techniques are needed.

2.2 Bayesian Inference by MCMC for CSSR Models

The parameters of the model are $\theta = (\beta_0, \ldots, \beta_2, \lambda_1, \lambda_2, \sigma_y^2, \sigma_x^2, \sigma_z^2)$. Assuming block-wise independence, the prior distribution is given by

$$p(\theta) = \mathcal{N}[\beta \mid \beta_*, H_*] \prod_j^2 N[\lambda_j \mid \lambda_{j*}, \tau_{j*}^2] \prod_j^3 Ga[\sigma_j^2 \mid \sigma_{j*}^2 n_{j*}/2, n_{j*}/2].$$

We adopt the convention that all parameters with a star are known hyper-parameters of the prior distribution and those with ** are known hyper-parameters of the posterior distribution. Let $\mathcal{D} = \{y, x, z\}$ denote the observed data, then the likelihood function is

$$l(ln\ y \mid \mathcal{D}, \theta) = N[ln\ y \mid X\beta, \sigma_\epsilon^2 I_n] N[ln\ x \mid \mu_x, \sigma_x^2 I_n] N[ln\ z \mid \mu_z, \sigma_z^2 I_n] * J \tag{3}$$

where J is the appropriate Jacobian of the model. For the multiplicative model this is $J = 1 - \frac{\beta_1}{1-\beta_1}\frac{\beta_2}{1-\beta_2}$. From the posterior distribution for θ, which is proportional to

$$p(\theta \mid \mathcal{D}) \propto l(ln\ y \mid x, z, \theta) p(\theta) \tag{4}$$

we can work out the posterior simulator for θ by MCMC using the following full conditional distributions (fcd) for the posteriors:

1. **The fcd for** β is given by

$$p(\beta \mid y, \ldots) \propto N[\beta \mid \beta_*, H_*] l(ln\ y \mid x, z, \theta) \\ N[ln\ x \mid \mu_x, \sigma_x^2 I_n] N[ln\ z \mid \mu_z, \sigma_z^2 I_n].$$

The last two components contain also β's because of the SPD constraints. Since this is not a known density we have to employ a Metropolis step, e.g. a random walk chain for the proposal β^{new}

$$\beta^{new} = \beta^{old} + N[0, c_\beta I_3]$$

where β^{old} is the previous generated value and c_β is a tuning constant for the variance. The acceptance probability involves the whole posterior density in (4) and is

$$\alpha(\beta^{old}, \beta^{new}) = min\left(\frac{p(\beta^{new})}{p(\beta^{old})}, 1\right).$$

2. **The fcd for** $\lambda_j, j = 1, 2$ (the average utility levels) is given in the 'usual' way, as

$$p(\lambda_1 \mid y, \ldots) \propto N[\lambda_1 \mid \lambda_{1*}, \tau_{1*}] N[ln\ x \mid \mu_x, \sigma_x^2 I_n], \tag{5}$$

$$p(\lambda_2 \mid y, \ldots) \propto N[\lambda_2 \mid \lambda_{2*}, \tau_{2*}] N[ln\ z \mid \mu_z, \sigma_z^2 I_n], \tag{6}$$

where the second normal kernels can be viewed as sort of likelihood function. Again we need a Metropolis step:

$$\lambda_j^{new} = \lambda_j^{old} + N[0, c_{\lambda,j}]$$

where $c_{\lambda,j}$ is a small proposal variance. The acceptance probability is

$$\alpha(\lambda_j^{old}, \lambda_j^{new}) = min\left(\frac{p(\lambda_j^{new})}{p(\lambda_j^{old})}, 1\right),$$

where $p(.)$ is the corresponding fcd from (5) or (6).
A direct derivation shows that the pdf is a conjugate normal density:

$$\tau_{1**}^{-2} = \tau_{1*}^{-2} + \sigma_x^{-2}(1 - \beta_1)^2$$

and

$$\lambda_{1**} = \tau_{1**}^2 [\tau_{1*}^{-2} \lambda_{1*} + \sigma_x^{-2}(1 - \beta_1)^2 \lambda_1].$$

3. **The fcd for** σ_j, $j \in y, x, z$ are given by

$$p(\sigma_j^2 \mid y, \ldots) \propto Ga[\sigma_j^2 \mid \sigma_{j**}^2 n_{j**}/2, n_{j**}/2]$$

with

$$n_{j**} = n_{j*} + n$$

and

$$n_{j**}\sigma_{j**}^2 = n_{j*}\sigma_{j*}^2 + e_j' e_j$$

where $e_j = lnj - \mu_j$ being the current residuals of the three regression equations and for $j \in y, x, z$.

Finally, MCMC in the CSSR model takes the following steps:

1. Starting values: set $\beta = \beta_{OLS}$ and $\lambda = 0$.
2. Draw σ_y^{-2} from $\Gamma[\sigma_y^{-2} \mid s_{y**}^2, n_{y**}]$.
3. Draw σ_x^{-2} from $\Gamma[\sigma_x^{-2} \mid s_{x**}^2, n_{x**}]$.
4. Draw σ_z^{-2} from $\Gamma[\sigma_z^{-2} \mid s_{z**}^2, n_{z**}]$.
5. Draw λ_j from $p(\lambda_j \mid \lambda_{j**}, \sigma_{j**}^{-2})$.
6. Draw β from $p[\beta \mid \mathbf{b}_*, \mathbf{H}_*]\, l(\theta \mid y)$.
7. Repeat until convergence.

3 A Spatial Auto-Regressive Extension to CSSR Models

Since the seminal work by Anselin (1988), spatial interactions have become an important tool in econometrics. Spatial applications have become popular in applied sciences, like in economics and also social sciences.

3.1 Spatial Lags

Consider a regression model where the dependent variable $\mathbf{y} = (y_1, \ldots, y_n)'$ is not independently observed but can be spatially correlated given the $n \times K$ matrix of independent observations $\mathbf{X}$. To model the spatial dependence we have to know (or specify) a spatial weight matrix $\mathbf{W}$ which has 3 properties: (1) All entries are positive, (2) the main diagonal elements are zero, and (3) all row sums are 1. Such a weight matrix could be a distance matrix if the y's are observed at geographical locations, it could be the first nearest neighbor only, but also a set of all contiguous neighbors.

This allows to specify a spatial lag variable of the dependent variable $\tilde{\mathbf{y}} = \mathbf{W}\mathbf{y}$. Each element of $\tilde{\mathbf{y}}$, i.e., $\tilde{y}_j = \mathbf{w}_j \mathbf{y}$ is a new "neighborhood observation", which summarizes the influence of the neighbors in form of a weighted average of the dependent variable and the j^{th} row vector $\mathbf{w}_j$.

Therefore we can formulate a 'structural' form of the spatial SAR model in the following form:

$$\mathbf{y} = \mathbf{X}\beta + \rho\mathbf{W}\mathbf{y} + \epsilon, \quad \epsilon \sim N\left[\mathbf{0}, \sigma^2\mathbf{I}_n\right], \tag{7}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix and ρ is the spatial correlation parameter. If ρ is zero then the model reduces to a simple regression model with independent errors.

Additionally, a reduced form is available by shifting all dependent variables on the left hand side:

$$\mathbf{z} = \mathbf{y} + \rho\mathbf{W}\mathbf{y} = \mathbf{X}\beta + \epsilon, \quad \epsilon \sim N\left[\mathbf{0}, \sigma^2\mathbf{I}_n\right].$$

Using the spread matrix $\mathbf{R}$ and its inverse $\mathbf{R}^{-1} = (\mathbf{I}_n - \rho\mathbf{W})^{-1}$, the reduced form

$$\mathbf{y} \sim N\left[\mathbf{R}^{-1}\mathbf{X}\beta, \sigma^2(\mathbf{R}'\mathbf{R})^{-1}\right],$$

is available since $Var(\mathbf{R}\epsilon) = \sigma^2\mathbf{R}\mathbf{R}'$. The prior distribution for the parameter $\theta = (\beta, \sigma^{-2}, \rho)$ is given by the product of (independent) blocks of normal and gamma distributions:

$$\begin{aligned} p(\beta, \sigma^{-2}, \rho) &= p(\beta) \cdot p(\sigma^{-2}) \cdot U[\rho \mid -1, 1] \\ &= N[\beta \mid \mathbf{b}_*, \mathbf{H}_*] \cdot \Gamma[\sigma^{-2} \mid s_*^2, n_*]\frac{1}{2}, \end{aligned}$$

where $U[-1, 1]$ stands for a uniform distribution in the interval $(-1, 1)$. Because of restrictions, the interval of feasible ρ's depends on λ_{min} and λ_{max}, the minimum and maximum eigenvalue of $\mathbf{W}$. It can be shown $\lambda_{min}^{-1} < 0$ and $\lambda_{max}^{-1} > 0$ and therefore ρ_t must lie between these bounds. Therefore, we restrict the prior space of ρ to the interval $(\lambda_{min}^{-1}, \lambda_{max}^{-1})$.

The joint distribution for $\mathbf{y}$ and the parameter $\theta = (\beta, \sigma^2, \rho)$ is

$$p(\beta, \sigma^{-2}, \rho, \mathbf{y}) \propto N\left[\mathbf{y} \mid \mathbf{X}\beta, \sigma^2\right] \cdot N[\beta \mid \mathbf{b}_*, \mathbf{H}_*] \cdot \Gamma[\sigma^{-2} \mid s_*^2, n_*].$$

3.2 The CSSR-SAR Model

The CSSR-SAR model is the CSSR model as in (2) with a spatial lag:

$$\begin{aligned} ln\ y &\sim N[\mu_y = \rho W ln\ y + X\beta, \sigma_y^2 I_n] \quad \text{or} \\ ln\ y &= \rho W ln\ y + \beta_0 + \beta_1 ln\ x + \beta_2 ln\ z + \epsilon \end{aligned}$$

with $\epsilon \sim N(0, \sigma_y^2 I_n)$. The partial derivative restrictions for the two regressors stay the same

$$ln\ x \sim N(\mu_x, \sigma_x^2 I_n) \qquad \text{and } ln\ z \sim N(\mu_z, \sigma_z^2 I_n).$$

The parameter vector is $\theta = (\beta_0, \ldots, \beta_2, \lambda_1, \lambda_2, \sigma_y^2, \sigma_x^2, \sigma_z^2, \rho)$ and includes the spatial ρ. The prior is – proportionally – the same (constant) since we assume uniform prior for ρ: $U[\rho \mid -1, 1] = 0.5$. The reduced form of the model is

$$ln\ y \sim \mathcal{N}\left[\mathbf{R}^{-1}\mathbf{X}\beta, \sigma^2(\mathbf{R}'\mathbf{R})^{-1}\right],$$

because $Var(\mathbf{R}\epsilon) = \sigma^2\mathbf{R}\mathbf{R}'$. This expression will now be used in the likelihood function

$$l(ln\ y \mid \mathcal{D}, \theta) = N[ln\ y \mid X\beta, \sigma_\epsilon^2(R'R)^{-1}] \tag{8}$$

$$N[ln\ x \mid \mu_x, \sigma_x^2 I_n]N[ln\ z \mid \mu_z, \sigma_z^2 I_n] * J \tag{9}$$

where J is the Jacobian of the model as before. For MCMC we can use the fcd results of the previous section. We just have to specify the additional fcd:

$$p(\rho \mid y, \ldots) \propto |\mathbf{I}_n - \rho\mathbf{W} \mid \exp\left(-\frac{\epsilon_\rho'\epsilon_\rho}{2\sigma_y^2}\right)$$

where the residuals of the spatial regression are

$$\epsilon_\rho = ln\ y - \mathbf{X}\beta - \rho W ln\ y.$$

Again, an additional Metropolis step is needed. We use

$$\rho^{new} = \rho^{old} + c_\rho\phi, \quad \phi \sim \mathcal{N}[0, 1]$$

where the scalar c_ρ is a tuning parameter and ρ^{old} the parameter of the previous value. The acceptance probability is

$$\alpha(\rho^{old}, \rho^{new}) = min\left(\frac{p(\rho^{new})}{p(\rho^{old})}, 1\right),$$

where p is the full conditional distribution above. Finally, the MCMC procedure has just to add one more draw for the ρ parameter:

1. Starting values: set $\rho = 0, \beta = \beta_{OLS}$ and $\lambda = 0$
2. Draw σ_y^{-2} from $\Gamma[\sigma_y^{-2} \mid s_{y**}^2, n_{y**}]$
3. Draw σ_x^{-2} from $\Gamma[\sigma_x^{-2} \mid s_{x**}^2, n_{x**}]$
4. Draw σ_z^{-2} from $\Gamma[\sigma_z^{-2} \mid s_{z**}^2, n_{z**}]$
5. Draw λ_j from $p(\lambda_j \mid \lambda_{j**}, \sigma_{j**}^{-2})$
6. Draw β from $p[\beta \mid \mathbf{b}_*, \mathbf{H}_*]\, l(\theta \mid y)$
7. Draw ρ using $p(\rho \mid \beta, \sigma_y^{-2})$
8. Repeat until convergence.

4 Empirical Test

The above models have been implemented using R and tested using synthetic and empirical data. Firstly, synthetic data with respect to a true model have been drawn using a bivariate sales response function where the x and z regressors were generated according to the above described SPD constraints. Repeatedly, a sample with size $n = 20$ was drawn and used as input for the MCMC algorithm, each with 1,000 repetitions. The average acceptance rate was 55.3%, the parameters of the true model could be adequately reproduced, indicating that the modeling framework and the MCMC algorithm work even for small samples.

Secondly, actual regional data from German pharma marketing were used for testing. For a specific brand and $n = 1{,}900$ (standard) regions, sales and marketing efforts data were available for Germany: brand and category prescriptions at the aggregated pharmacy wholesale level from a market research institute as well as number of visits of physicians by the company's salespersons. Additionally, data on further regional marketing activities could be used. The regions differ with respect to the neighboring regions, purchasing power, population density, number of types of physicians as well as category sales and brand shares. Assumed are differences with respect to the sales elasticities by, e.g., population density (urban vs. rural regions), average number of patients per physician or purchasing power per inhabitant. These differences were assumed to have an influence on the effectiveness of the marketing expenditures.

For testing the above CSSR model, sales figures (for one brand) and corresponding regressors (visits of two types of relevant physicians) were selected. Since the visits are very costly, one could expect that the company allocates them according to the assumed equal marginal returns rule. The analysis showed, that the CSSR models could be calibrated by the MCMC algorithms. However, it also showed by high λ-values, that the expected optimal allocation rule wasn't obeyed across all regions.

5 Conclusions and Outlook

New sales response functions were proposed, the Cross Sectional Sales Response (CSSR) model and its Spatial Auto Regressive (SAR) extension taking the neighborhood structure of the observations into account. They make use of Stochastic Partial Derivative (SPD) constraints and of MCMC algorithms for improving the parameter estimation. First tests with simulated and real data show promising results. However, more tests with simulated and real data are needed for assessing the benefits of these sales response models over the traditional ones in more detail.

References

Anselin, L. (1988). Spatial econometrics. In: B. H. Baltagi (Ed.), *A companion to theoretical econometrics* (pp. 310–330). Oxford: Blackwell Publishing Ltd AD.

Kao, L.-J., Chiu, C.-C., Gilbride, T. J., Otter, T., & Allenby, G. M. (2005). Evaluating the effectiveness of marketing expenditures. *Working Paper*, Ohio State University, Fisher College of Business.

Lerer, L. (2002). Pharmaceutical marketing segmentation in the age of the internet. *International Journal of Medical Marketing, 2*(2), 159–166.

Queitsch, M., & Baier, D. (2005). e-Detailing in der Pharmaindustrie: Herausforderungen, technologien und integration. *Pharma Marketing Journal, 4*, 134–138.

Rossi, P. E., Allenby, G. M., & McCulloch, R. (2005). *Bayesian statistics and marketing.* New York: John Wiley and Sons.

A Demand Learning Data Based Approach to Optimize Revenues of a Retail Chain

Wolfgang Gaul and Abdolhadi Darzian Azizi

Abstract We consider the problem of selling a perishable product via a retail chain during a restricted time interval. Based on data describing buying behavior in the outlets of the retail chain in the first part of a specified selling season, we provide an optimal policy for reallocating the remaining inventory of the product and setting prices in the remaining part of the selling season in order to optimize total expected revenues. We use a Bayesian update approach in which the retail chain learns about customers' demand patterns in each outlet during earlier periods and decides how to allocate the remaining stock and adjust prices w.r.t customer arrival rates and willingness to pay. An illustrative example is used to describe our demand learning approach.

1 Introduction

Retailers who are selling perishable products usually face problems such as *demand uncertainty* and *supply inflexibility* to optimize revenues. Demand uncertainty refers to the lack of information about how successful a product will be in the market and supply inflexibility describes a situation in which one has no opportunity to replenish inventory during the selling season. In such cases, retailers have to apply effective pricing policies to reduce the costs of mismatching between demand and supply. For example, M&S lost £150 million due to failures in matching supply with demand in 1998–1999 (Christopher and Towill 2002).

For retail chains that operate in different markets (w.r.t outlets, sales channels, etc.) the determination of successful pricing policies is even more difficult because demand patterns could vary among markets. In this paper, we consider the problem of selling a perishable product in several outlets over a finite horizon of time, provide

W. Gaul (✉)
Institut für Entscheidungstheorie und Unternehmensforschung, Karlsruhe University, Karlsruhe, Germany
e-mail: wolfgang.gaul@wiwi.uni-karlsruhe.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_75,

a Bayesian update approach in which we learn about customers' demand patterns in each outlet during earlier periods of the sales season, and update the parameters of the demand distributions by using these observations to determine an optimal selling policy based on (1) inventory, (2) time, and (3) demand patterns in order to maximize the total expected revenues.

Problems of this kind are considered in the literature on revenue management in which optimal pricing policies are determined based on inventory on-hand and remaining time until the end of the selling season (e.g., Bitran and Mondschein 1993; Gallego and van Ryzin 1994; Bitran and Caldentey 2003). In practice, however, there are many situations where sellers do not have full knowledge concerning demand patterns. Thus, demand learning is an effective approach to resolve demand uncertainty in which a decision maker improves his knowledge about the real situation based on the observed demand during the sales season and updates forecasts of future demand (e.g., Lazear 1986; Jorgensen et al. 1999; Burnetas and Smith 2001; Petruzzi and Dada 2002; Lin 2005).

Studies on retail chain management have considered the situation of clearance (markdown) pricing policies without demand learning (e.g., Smith and Achabal 1998; Bitran et al. 1998) and also the determination of an optimal pricing policy based on demand learning (e.g, Elmaghraby and Keskinocak 2003). In this paper, we develop a Bayesian learning approach to revise the parameters of the future demand distribution for each outlet and provide an algorithm to determine the optimal selling policy including an optimal price and a reallocation w.r.t the remaining inventory at the beginning of a future period of time.

The organization of this paper is as follows: After describing our model and main assumptions in Chap. 2, we provide an optimal selling policy on the basis of a Bayesian learning approach. In Chap. 3, we present a numerical study to show how our approach is able to tackle the underlying situation. Finally, we summarise our work in Chap. 4 and suggest future applications for our demand learning approach.

2 Model Description

We consider a retailer who orders a fixed amount of a perishable product before the selling season $[0, T]$ which is divided into T equal periods $[t-1, t]$, $t = 1, \ldots, T$. The retailer has J outlets across a monopolistic market and charges the same price for the product in all outlets during a given period $[0, t]$. At decision point t, we assume that the retailer updates the price and reallocates the remaining inventory for the rest of the selling season. We also assume that there exists a predetermined set of prices P_t for the product after period t.

As we can see in Fig. 1, the retailer orders the initial inventory of the product, q_0, before the sales season. At time zero, he sets the initial price at p_0 and determines the initial allocation of inventory to each outlet, q_{j0}, $j = 1, \ldots, J$. Then, he monitors the arrivals of customers $n_j = \{n_{j1}, \ldots, n_{jt}\}$ and the sales $s_j = \{s_{j1}, \ldots, s_{jt}\}$ w.r.t outlet j in the first t periods. At the end of period t, he uses the observations

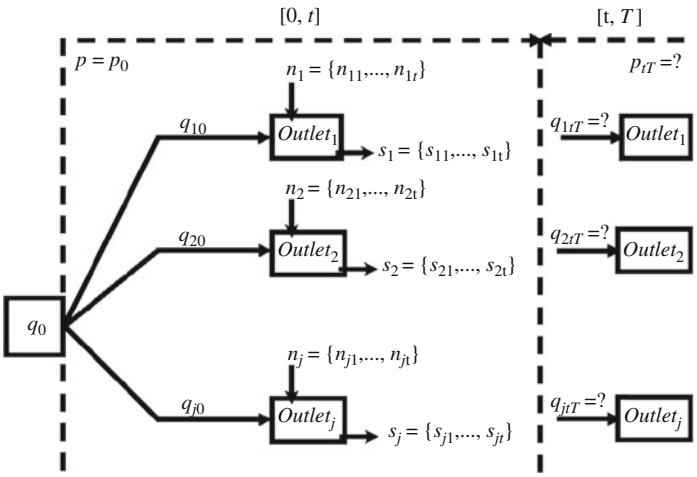


Fig. 1 A scheme of a retail chain operation in several outlets

collected so far to determine an optimal selling policy for the remaining time $(t, T]$. Here, two questions arise:

1. At which price should the product be sold during the remaining time?
2. How many units of the remaining inventory should be reallocated among the outlets for the remaining time?

To determine an optimal selling policy by using data from earlier periods, we need a model to forecast future demand for the product in each outlet. We assume that demand for the product is the result of the customers' arrival rates and reservation prices. In terms of the arrival rates, we assume that customers arrive at outlet j, $j = 1, \ldots, J$, according to a Poisson process with mean λ_j that does not depend on the posted price, is unknown to the retailer, but follows a gamma distribution with density function

$$f(\lambda_j; a_j, b_j) = \frac{b_j e^{-b_j \lambda_j} (b_j \lambda_j)^{a_j - 1}}{\Gamma(a_j)}, \lambda_j \geq 0. \tag{1}$$

Under this assumption, the probability distribution of the total arrivals $\bar{N}_{jt}$ at outlet j during time-interval $[0,t]$ is

$$\begin{aligned} P(\bar{N}_{jt} = n) &= \int_0^\infty P(\bar{N}_{jt} = n|\lambda_j) f(\lambda_j; a_j, b_j) d\lambda_j \\ &= \frac{\Gamma(a_j + n)}{n!\Gamma(a_j)} \left(\frac{b_j}{b_j + t}\right)^{a_j} \left(\frac{t}{b_j + t}\right)^n, n = 0, 1, \ldots, \end{aligned} \tag{2}$$

(where $\Gamma(k) = (k-1)! = \int_0^\infty x^{k-1}e^{-x}dx$ is the gamma function), i.e., the probability of observing $\bar{N}_{jt}$ arrivals is given by a negative binomial distribution that performs well in retailing (e.g., Agrawal and Smith 1996). If the number of arrivals at outlet j during the t earlier periods is equal to $\bar{n}_{jt} = \sum_{k=1}^{k=t} n_{jk}$, the posterior distribution of the customers' arrival rate will be again a gamma distribution, now with parameters $(a_j + \bar{n}_{jt}, b_j + t)$ (e.g., DeGroot and Schervish 2002):

$$g(\lambda_j) = \frac{(b_j+t)e^{(-b_j+t)\lambda_j}((b_j+t)\lambda_j)^{a_j+\bar{n}_{jt}-1}}{\Gamma(a_j+\bar{n}_{jt})}, \lambda_j \geq 0. \tag{3}$$

Using (3), the probability distribution of the arrivals at outlet j during periods $t+1,\ldots,T$ will be

$$P(\bar{N}_{jT}-\bar{n}_{jt}=n)=\frac{\Gamma(a_j+\bar{n}_{jt}+n)}{n!\Gamma(a_j+\bar{n}_{jt})}\left(\frac{b_j+t}{b_j+T}\right)^{a_j+\bar{n}_{jt}}\left(\frac{T-t}{b_j+T}\right)^{n}, n=0,1,\ldots, \tag{4}$$

where $(\bar{N}_{jT}-\bar{n}_{jt})$ is the number of arriving customers at outlet j during time $(t,T]$, i.e., we have again a negative binomial distribution.

In terms of the customers' reservation prices, we assume that each arriving customer at outlet j has a willingness to pay equal to v_j that is a random variable with a continuous cumulative distribution function $F_j^{RP}(.)$. The retailer knows that an arriving customer will probably purchase one unit of the product if the posted price is not larger than his reservation price. That is, when the retailer sets the price at p, an arriving customer will purchase the product if $v_j \geq p$ (with probability $\bar{F}_j^{RP}(p) = 1 - F_j^{RP}(p)$). In the literature on revenue management exponential and Weibull distributions are usually used to represent the probability density function of the customers' reservation price (e.g., Bitran and Wadhwa 1996; Bitran and Mondschein 1997). Therefore, the expected demand for the product in outlet j during periods $t+1,\ldots,T$, at price p_{tT}, is given by

$$\bar{D}_{jtT}(p_{tT}) = \sum_{n=1}^{\infty} nP(\bar{N}_{jtT}=n)(1-F_j^{RP}(p_{tT})), \tag{5}$$

where $\bar{N}_{jtT} = \bar{N}_{jT} - \bar{n}_{jt}$ and $P(\bar{N}_{jtT}=n)$ is the probability that n customers arrive at outlet j during the remaining periods [see (4)]. It is worth noting that our Bayesian approach allows to consider the situation where both the customers' arrival rate and the customers' willingness to pay can be different among the outlets. In this setting, the optimal revenue problem can be formulated as:

$$R = \max_{p_{tT}\in P} \sum_{j=1}^{J} \overbrace{p_{tT}\bar{S}_{jtT}(p_{tT},q_{jtT})}^{Expected\ Revenue\ for\ outlet\ j} \tag{6}$$

$$\bar{S}_{jtT}(p_{tT},q_{jtT}) = q_{jtT}P(\bar{D}_{jtT}(p_{tT}) \geq q_{jtT})$$

$$s.t. \qquad \sum_{j=1}^{J} q_{jtT} \leq Q_t$$
$$p_{tT} \in P_t, q_{jtT} \in \{0, 1, \ldots, Q_t\},$$

where

$\bar{S}_{jtT}(.)$: Expected sales in outlet j during the periods $t+1, \ldots, T$;
$\bar{D}_{jtT}(.)$: Expected demand in outlet j during the periods $t+1, \ldots, T$;
p_{tT} : Posted price over the periods $t+1, \ldots, T$;
P_t : The set of prices for the product after period t;
q_{jtT} : Number of units allocated to outlet j for the periods $t+1, \ldots, T$;
Q_t : Quantity of the remaining units at the end of period t.

The optimal expected revenues after period t are given by the maximum of the sum of the expected revenues in the outlets. For the non-linear optimization problem (6) we use a greedy algorithm that is based on the idea of allocating each remaining unit of the product to that outlet which has the largest marginal expected revenues. Five steps have to be carried out:

Step 1: Compute the probability distribution of the customers' arrivals at each outlet during the next periods.
Step 2: Compute the probability of demanding at least x unit(s) in each outlet during the next periods, $x = 1, \ldots, Q_t$.
Step 3: Set different prices and compute the expected revenues of selling x unit(s) and the marginal expected revenues of selling the x^{th} unit in each outlet, $x = 1, \ldots, Q_t$.
Step 4: Find the maximum total expected revenues w.r.t each price in P_t.
Step 5: Find that price which creates the maximum total expected revenues.

3 Numerical Study

Consider $q_0 = 33$, $J = 2$, $T = 5$, and $t = 3$, i.e., a retailer wants to sell 33 units of a perishable product in two outlets during a five periods selling season. The retailer sets the initial price at $p_0 = 30$ and decides to update the price after 3 periods. Table 1 presents information w.r.t the outlets such as the arrivals, sales, and the parameters (a_j, b_j) of the gamma distributions for the three first periods by which we are able to compute the parameters of the arrival distributions and the parameter of the customer's reservation price distribution for the remaining periods. The parameters (a_j, b_j) are given by

$$a_j = \mu_{n_j}^2 / \sigma_{n_j}^2, b_j = \mu_{n_j} / \sigma_{n_j}^2. \tag{7}$$

where μ_{n_j} is the mean of the arrivals during the first three periods in outlet j and $\sigma_{n_j}^2$ the corresponding variance. The parameter of the customer's reservation price

Table 1 Information w.r.t the outlets of the retail chain

Outlets	n_{j1}	n_{j2}	n_{j3}	$\bar{n}_{j3}$	$\bar{s}_j$	a_j	b_j	θ_j
1	17	15	16	48	10	384	24	0.052
2	11	11	8	30	8	50	5	0.044

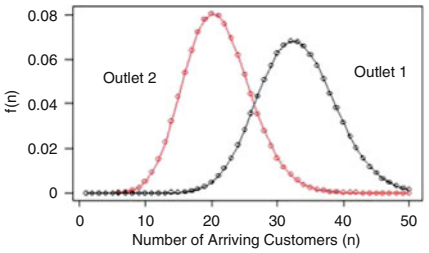


Fig. 2 The probability distributions of the arrivals at the outlets during the remaining periods

θ_j can be derived via

$$\bar{s}_j = \bar{n}_{j3}(1 - F_j^{RP}(p_0)) = \bar{n}_{j3}e^{-\theta_j p_0} \Rightarrow \theta_j = \frac{ln(\bar{n}_{j3}/\bar{s}_j)}{p_0}, \tag{8}$$

where

$\bar{s}_j$: Number of units sold in outlet j during periods $1, \ldots, 3$.

$\bar{n}_{j3}$: Number of customers that arrived at outlet j during periods $1, \ldots, 3$.

p_0 : Initial price of the product (the price during periods $1, \ldots, 3$).

θ_j : Parameter of the reservation price distribution in outlet j.

In the first step, we compute the probability distribution of the customers' arrivals at each outlet during the 4^{th} and 5^{th} periods via (4) and data from Table 1. Figure 2 shows the probability distributions of the arrivals at the outlets during the remaining periods.

In the second step, we compute the probability of demanding at least x unit(s) in each outlet during the next periods by

$$P(\bar{D}_{jtT} \geq x) = 1 - F_j^A(n)_{n \equiv min\{n: ne^{-\theta_{jt} p_{tT}} = x\}}, x = 1, \ldots, Q_t, \tag{9}$$

where $F_j^A(n)$ is the cumulative distribution function of the customers' arrivals at outlet j based on (4)(see Fig. 3).

In the third step, we consider different prices $p_{tT} \in P_t$ and compute the expected revenues (ER) of selling $x = 1, \ldots, Q_t$ unit(s) and the marginal expected revenues (MR) of selling the x^{th} unit in each outlet as follows:

$$ER_{jtT}(x) = ER_{jtT}(x-1) + P(\bar{D}_{jtT} \geq x)p_{tT}, ER_{jtT}(0) = 0. \tag{10}$$

$$MR_{jtT}(x) = ER_{jtT}(x) - ER_{jtT}(x-1) = P(\bar{D}_{jtT} \geq x)p_{tT}. \tag{11}$$

Figures 4 and 5 show the results of this step.

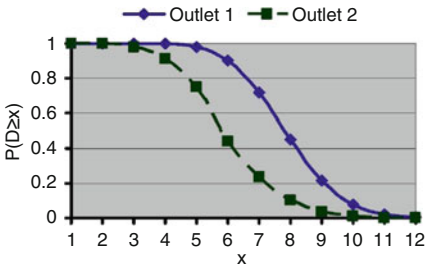


Fig. 3 The probabilities of demanding at least x unit(s) in the outlets during the remaining periods.

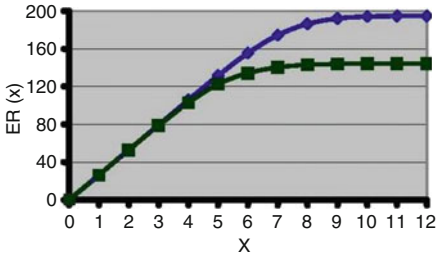


Fig. 4 The expected revenues of selling x units in the outlets

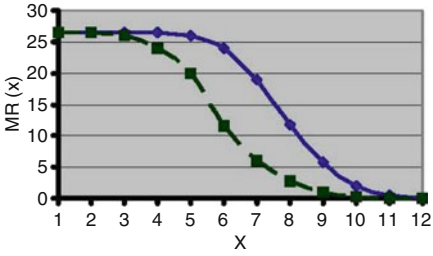


Fig. 5 The marginal revenues of selling the x^{th} unit in the outlets

In the fourth step, we find the maximum total expected revenue with respect to each price from P_3. And finally, we find that price which causes the maximum total expected revenues together with the corresponding inventory allocations to the outlets. Figure 6 shows the optimal price $p_{tT} = 26.4$ and the maximum total expected revenue $R = 326.5$. In this example, the optimal allocation of the 15 remaining units is $\{q_{135}, q_{235}\} = \{8, 7\}$.

4 Summary and Future Directions

In this paper, we studied the problem of selling a fixed amount of a perishable product via a retail chain during a sales season $[0, T]$ under demand uncertainty. We assumed that the demand for the product in each outlet can be determined with the help of the corresponding customers' arrival rate and their willingness to pay. Additionally, we assumed that the customers arrive at each outlet according to a Poisson process, whose mean follows a gamma distribution, and that an arriving customer

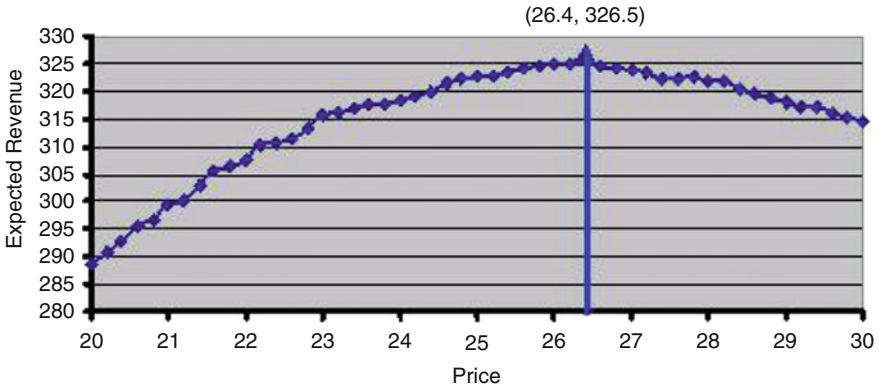


Fig. 6 The maximum total expected revenues of selling the remaining inventory $Q_t = 15$ as a function of price

purchases one unit of the product according to a "willingness-to-pay" distribution function. We developed a Bayesian approach by which the retailer is able to learn about the demand patterns in each outlet as time progresses and can determine the optimal price and the optimal allocation of the remaining inventory in order to maximize his total expected revenues. An "easy-to-understand" example was used to demonstrate how the overall procedure works. Of course, different specifications of the underlying model parameters and moving t situations can be considered. Additionally, several extensions of the here described situation are of interest for future studies, e.g., considering the case in which the retailer is allowed to sell the product at different prices in different outlets, allowing (optimal) replenishments during the selling season, or taking into account non-homogeneous arrivals.

References

Agrawal, N., & Smith, S. A. (1996). Estimating negative binomial demand for retail inventory management with unobservable lost sales. *Naval Research Logistics, 46*(6), 839–861.

Bitran, G. R., & Caldentey, R. (2003). An overview of pricing models for revenue management. *Manufacturing and Service Operations Management, 5*, 203–229.

Bitran, G. R., Caldentey, R., & Mondschein, S. V. (1998). Coordinating clearance markdown sales of seasonal products in retail chains. *Operations Research, 46*, 609–624.

Bitran, G. R., & Mondschein, S. V. (1993). Perishable product pricing: An application to the retail industry. M.I.T. Sloan School of Management, Working Paper.

Bitran, G. R., & Mondschein, S. V. (1997). Periodic pricing of seasonal products in retailing. *Management Science, 43*, 427–443.

Bitran, G. R., & Wadhwa, H. K. (1996). A methodology for demand learning with an application to the optimal pricing of seasonal products. M.I.T. Sloan School of Management, Working Paper.

Burnetas, A. N., & Smith, C. E. (2001). Adaptive ordering and pricing for perishable products. *Operations Research, 48*, 436–443.

Christopher, M., & Towill, D. R. (2002). Developing market specific supply chain strategies. *The International Journal on Logistics Management, 13*, 1–14.

DeGroot, M. H., & Schervish, M. J. (2002). *Probability and Statistics* (3rd ed.). Reading, MA: Addison-Wesley.

Elmaghraby, W., & Keskinocak, P. (2003). Dynamic pricing in the presence of inventory considerations: Research review, current practices, and future directions. *Operations Research, 49*, 1287–1309.

Gallego, G., & van Ryzin, G. J. (1994). Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science, 40*, 999–1020.

Jorgensen, S., Kort, P. M., & Zaccour, G. (1999). Production, inventory, and pricing under cost and demand learning effect. *European Journal of Operational Research, 17*, 382–395.

Lazear, E. (1986). Retail pricing and clearance sales. *The American Economic Review, 76*, 14–32.

Lin, K. Y. (2005). Dynamic pricing with real-time demand learning. *European Journal of Operational Research, 174*, 522–538.

Petruzzi, N. C., & Dada, M. (2002). Dynamic pricing and inventory control with learning. *Naval Research Logistics, 49*, 303–325.

Smith, S. A., & Achabal, M. (1998). Clearance pricing and inventory policies for retail chains. *Management Science, 44*, 285–300.

Missing Values and the Consistency Problem Concerning AHP Data

Wolfgang Gaul and Dominic Gastes

Abstract For AHP (Analytic Hierarchy Process) applications a paired comparison data matrix with special structure is needed. The limited ability of human respondents to provide paired comparison data that fulfil a consistency constraint and absent knowledge to judge all pairs of items under consideration can often result in data for which a best approximating complete AHP matrix has still to be found.

We compare existing methodologies for the problem and describe an own approach to handle this situation. An illustrative example is used to explain our findings.

1 Introduction

The Analytic Hierarchy Process (AHP) is a frequently used method in the field of multicriteria decision making and has helped to support decision makers in many different application areas. For recent overviews concerning applications see Ho (2008) and Vaidya and Kumar (2006). Typically a main objective of decision makers using the AHP is to calculate priority weights for hierarchically organized elements on different levels. These elements can be objectives, subobjectives, tangible or intangible criteria or attributes, and, at the bottom level, alternatives, that are the basic items that one wants to value with the help of hierarchically ordered subproblems. Within a sub-problem paired comparisons of the elements belonging to the underlying level have to be carried out. Thus, a multi criteria decision problem is divided into h ($h = 1, \ldots, H$) sub-problems. Each sub-problem h contains n^h elements, which have to be pairwise compared to each other w.r.t. an element of the next higher level. By pairwise comparisons between all those elements linked within sub-problem h, a pairwise comparison matrix A^h can be constructed. In the following the index of sub-problem h will be omitted, because the further investigations do not focus on hierarchical structures or aggregation procedures but on the pairwise comparison matrix in an underlying sub-problem.

D. Gastes (✉)
Institute of Decision Theory and Operations Research, University of Karlsruhe,
Kaiserstrasse 12, 76131 Karlsruhe, Germany
e-mail: dominic.gastes@wiwi.uni-karlsruhe.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_76,

When decision makers are asked to compare two objects i and j, the discussion about the limitations of human capacity for processing information has to be considered (see Miller 1956; Saaty and Ozdemir 2003). This is one reason, why the AHP often uses the so-called Saaty scale, which contains the numbers $1, 2, \ldots, 8, 9$ and their reciprocals. When comparing two elements a value of 1 is assigned to the verbal statement "element i and element j are equally important", a value of 3 is assigned to the verbal expression "element i is preferred to element j" and a value of 9 is assigned to "element i is extremely preferred (or absolute dominant) w.r.t. element j". Intermediate values can be used for more detailed assessments and reciprocals for vice versa evaluations (see Saaty 1980). Thus, one gets the matrix

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \tag{1}$$

under the assumptions

$$a_{ij} \in \left\{ \frac{1}{9}, \frac{1}{8}, \ldots, 1, \ldots, 8, 9 \right\} \text{ (Saaty scale)} \tag{2}$$

$$a_{ij} \cdot a_{ji} = 1 \ \forall \ i, j \text{ (reciprocity)} \tag{3}$$

$$a_{ij} = a_{ik} \cdot a_{kj} \ \forall \ i, j, k \text{ (consistency)} \tag{4}$$

If assumptions (2), (3), (4) hold, the priority weights of the n elements can be calculated by solving the eigenvalue problem:

$$A\mathbf{w} = \lambda_{max}\mathbf{w} \tag{5}$$

The normalized eigenvector $\mathbf{w} = (w_1, \ldots, w_n)'$, which is assigned to the principal eigenvalue λ_{max} of A, represents the priority weights of the decision maker for the n elements. In addition λ_{max} is equal to the number of elements n and for all entries $a_{ij} = \frac{w_i}{w_j}$ is valid (see Saaty 1980). In applications assumption (4) is often violated. Thus, a consistency index (CI) was developed to measure the degree of violation of the consistency assumption of matrix A (see Saaty 2003), which has the form

$$CI = \frac{\lambda_{max} - n}{n - 1}. \tag{6}$$

To get a normalized measure for different sized matrices A the CI-value is related to the average random consistency index $RI(n)$ of randomly chosen reciprocal matrices of size $n \times n$ constructed via entries the values of which are taken from the Saaty scale [see (2)]. The so-called consistency ratio (CR) of matrix A is given as $CR = \frac{CI}{RI(n)}$, A is called "full consistent", if $CR = 0$. A is called "α consistent" if $0 < CR \leq \alpha$ and "α inconsistent" if $CR > \alpha$ for $\alpha \in (0, 1]$. Usually

$\alpha = 0,1$ is chosen. The limiting condition of $CR \leq 0,1$ was explained by statistical experiments (see Vargas 1982).

Full, or at least α consistency ($\alpha \leq 0,1$) is necessary, because **w** reflects cardinal preferences yielded by the ratios $\frac{w_i}{w_j}$, that approximate the entries of A. To preserve the strength of preferences, **w** has to be invariant under multiplication by a positive constant and also under hierarchic composition w.r.t. its own judgment matrix. These conditions for **w** are satisfied, if A is α consistent (see Saaty 2003).

α consistency instead of full consistency may be due to the limitations of the Saaty scale, inaccuracies caused by of a lack of concentration or uncertainties of the decision maker. Especially, if the number of objects, which have to be compared, is high, it is a challenging task for decision makers to provide all comparisons in a full or even α consistent way (see Decker et al. 2008 for a simulation study on automated detecting and debugging different kinds of erroneous statements in comparison matrices).

2 Consistency Adjustment Approaches

In applications comparison matrices are often neither full nor α consistent. In an ongoing study we conducted a paper and pencil AHP resulting in 13 comparison matrices of size 6×6 and 78 matrices of size 7×7. 58% of the 91 matrices were not 0.1 consistent (a similar percentage can be found in a study containing 84 comparison matrices Lin et al. 2008). If a comparison matrix is not at least 0.1 consistent, it should be adjusted (see Saaty 2003), for which several methods can be used. One can distinguish between manual and automated consistency adjustment approaches.

2.1 Manual Consistency Adjustment Approaches

A way of dealing with inconsistency is to control consistency during the data collection processes. A kind of tutoring system can be used, which guides the decision maker by showing a range in which a comparison value has to be, to stay within a given consistency limit (see Ishizaka and Lusti 2004).

Another way of manual consistency adjustments is the reassessment approach. It is used in many AHP software products. If, after data collection, the comparison matrix A is not sufficiently consistent, the software selects the "most inconsistent entry" $\hat{a}_{ij}$ and the decision maker is asked to reconsider and probably reassess his evaluation for this entry. This procedure can be repeated until A is at least α consistent.

One advantage of manual approaches for consistency adjustments is that the decision maker is forced to reconsider inconsistent preferences and that inaccuracies or mix-ups can be detected and corrected. A big disadvantage is that a manual reassessment approach is very time-consuming and that a decision maker, who is

concentrated on evaluating elements in his imagination of consistency, may – now – answer not according to his conviction only to respond to the consistency standards that he is urged to fulfil.

2.2 Automated Consistency Adjustment Approaches

Automated consistency adjustment methods start with a given comparison matrix A, which is not sufficiently consistent, and try to find α or full consistent matrices without asking the decision maker for activities.

2.2.1 Automated Expert-Choice Method (AEM)

Given an inconsistent matrix A with eigenvalue λ_{max} and corresponding eigenvector $\mathbf{w}$, a method to adjust the consistency without reassessment is to select that entry a_{ij} with the largest corresponding error term $\epsilon_{ij} = a_{ij}\frac{w_j}{w_i}$, set a_{ij} and the reciprocal a_{ji} to zero, and calculate a more consistent value for a_{ij} with the help of a weight vector $\tilde{\mathbf{w}}$ of a modified matrix $\tilde{A}$ with $\tilde{a}_{ii} = \tilde{a}_{jj} = 2$ and $\tilde{a}_{ij} = \tilde{a}_{ji} = 0$. The entry a_{ij} in the original matrix A is replaced by $a_{ij} = \frac{\tilde{w}_i}{\tilde{w}_j}$ and $a_{ji} = \frac{\tilde{w}_j}{\tilde{w}_i}$ (see Harker 1987). We call this approach "Automated Expert-Choice Method" (AEM), because the idea of selecting the a_{ij} with the largest corresponding ϵ_{ij} as first entry to be adjusted is also used by the Expert Choice AHP software (see Saaty 2003).

2.2.2 Iterative Eigenvalue Improvement Method (IEM)

A different approach was suggested in Zeshui and Cuiping (1999). The authors suggested an algorithm, which modifies a comparison matrix A stepwise, so that the corresponding maximum eigenvalue is changed with each iteration r. The formula is $a_{ij}^{(r+1)} = (a_{ij}^{(r)})^{\rho}\left(\frac{\omega_i^{(r)}}{\omega_j^{(r)}}\right)^{1-\rho}$ with $0 < \rho < 1$ and $\boldsymbol{\omega}^{(r)}$ as corresponding normalized principal eigenvector of $A^{(r)}$. We call this approach "Iterative Eigenvector Improvement Method" (IEM).

2.2.3 Genetic Adjustment Method (GAM)

Our new approach is based on a genetic algorithm, which finds a "best fitting" comparison matrix X, w.r.t. the given matrix A, as solution of the following minimization problem:

$$\min : \left\{ \sum_{ij} (a_{ij} - x_{ij})^2 \right\}, \tag{7}$$

subject to

$$x_{ij} \in \left\{ \frac{1}{9}, \frac{1}{8}, \ldots, 1, \ldots, 8, 9 \right\} \qquad \text{(Saaty scale)} \tag{8}$$

$$x_{ij} \cdot x_{ji} = 1 \; \forall \; i, j \qquad \text{(reciprocity)} \tag{9}$$

$$CR \leq 0,1 \; (\alpha = 0,1 \text{ consistency}) \tag{10}$$

where $CR = CR(X)$.

This approach minimizes the difference between the given, but inconsistent evaluations A of the decision maker, and the adjusted and α consistent values of the solution X where $\alpha = 0,1$ is selected in this application. We call this approach "Genetic Adjustment Method" (GAM).

3 Comparison of Automated Consistency Adjustment Approaches

As described in the previous section, the AEM and the IEM produce stepwise adjusted matrices with decreasing consistency ratios. In every step at least two entries of the underlying matrix are changed. One should notice that the new entries of the adjusted matrices do not necessarily fulfil the Saaty scale assumption (2). On the one side the adjusted matrices may contain entries the values of which lie between the values of the Saaty scale, thus, they have no longer a linguistic representation w.r.t. the comparison situation that was formulated in terms of the Saaty scale. On the other side the matrices may contain entries the values of which are even outside of the bounds of the Saaty scale. Only the GAM approach provides a solution matrix with Saaty scale values and is directly comparable with the given matrix of the decision maker.

For comparisons between the different approaches 100 random comparison matrices A^k, $k = 1, \ldots, 100$, with corresponding consistency ratios between 0,5 and 1 with sizes of 7×7 as well as 8×8 were generated. Then the AEM, IEM, and our GAM solutions were calculated. Also rank vectors were used, because methods may lead to different weights, but can result in the same rank orders of the elements involved (see Zanakis et al. 1998).

3.1 Common Performance Measures

As one comparison measure between the inconsistent matrices (A^k) and the adjusted α consistent matrices ($A^k_{AEM}, A^k_{IEM}, A^k_{GAM}$) the mean squared matrix error (MSME) was used. For comparisons between the priority weight rank vectors ($\mathbf{w}^k, \mathbf{w}^k_{AEM}, \mathbf{w}^k_{IEM}, \mathbf{w}^k_{GAM}$) the mean squared priority weight vector error for ranks (MSPER) was computed as a performance measure. The larger the value of

this measure gets, the more do the rank vectors of the inconsistent matrices and the corresponding adjusted matrices differ. As an additional indicator for the relation between rank vectors, the Spearman correlation coefficient for ranks (SCR) was calculated (see Carmone et al. 1997; Zanakis et al. 1998). For matrices A,B these measures are:

$$MSME(A, B) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (a_{ij} - b_{ij})^2 \tag{11}$$

$$MSPER(\mathbf{w}_A, \mathbf{w}_B) = \frac{1}{n} \sum_{i=1}^{n} (w_{Ai} - w_{Bi})^2 \tag{12}$$

$$SRC(\mathbf{w}_A, \mathbf{w}_B) = \frac{6 \cdot \sum_{i=1}^{n} (w_{Ai} - w_{Bi})^2}{n(n^2-1)} \tag{13}$$

3.2 Results

In Table 1 it can be seen, that the average consistency ratios of the adjusted matrices are lower then $\alpha = 0,1$ and, thus, the adjusted matrices are α consistent. The GAM is in the 7×7 case as well as in the 8×8 case nearest to the given limit of α consistency ($CR \leq 0,1$). If lower α values are of interest this can easily be considered by a reformulation of constraint (10). The IEM can be influenced via the choice of the value for the parameter ρ. All following results are based on a parameter value of $\rho = 0,9$. Also simulations with values near to 0,9 were conducted, but did not differ much from the results achieved with the parameter value of $\rho = 0,9$ (see Zeshui and Cuiping 1999). Tables 2 and 3 show the average mean squared matrix errors between the corresponding matrices. It can be seen, that the GAM is nearest to the values given by the randomly selected A^k matrices. Tables 4 and 5 show the mean squared priority weight vector error for ranks. One can see, that the ranks derived from the inconsistent matrices and the ranks derived from the matrices adjusted with the IEM are most similar with respect to MSPER. The

Table 1 Overview of average consistency ratios

	$CR(A^k)$	$CR(A^k_{AEM})$	$CR(A^k_{IEM})$	$CR(A^k_{GAM})$
$A^k_{7\times7}$	0,8056	0,0841	0,0891	0,0981
$A^k_{8\times8}$	0,8340	0,0879	0,0886	0,0993

Table 2 Average MSME based on 100 7×7 matrices A^k

Average MSME (7×7)			
	A^k_{AEM}	A^k_{IEM}	A^k_{GAM}
A^k	6,1846	3,9784	2,8073
A^k_{AEM}		3,8486	3,5427
A^k_{IEM}			1,4855

Table 3 Average MSME based on 100 8×8 matrices A^k

Average MSME (8×8)			
	A^k_{AEM}	A^k_{IEM}	A^k_{GAM}
A^k	6,4438	4,4551	3,1912
A^k_{AEM}		3,8846	3,6792
A^k_{IEM}			1,6209

Table 4 Average MSPER based on 100 8×1 priority rank vectors

Average MSPER (7×1)			
	$\mathbf{w}^k_{AEM}$	$\mathbf{w}^k_{IEM}$	$\mathbf{w}^k_{GAM}$
$\mathbf{w}^k$	2,3257	0,2800	1,2857
$\mathbf{w}^k_{AEM}$		2,0600	2,6029
$\mathbf{w}^k_{IEM}$			1,0657

Table 5 Average MSPER based on 100 7×1 priority rank vectors

Average MSPER (8×1)			
	$\mathbf{w}^k_{AEM}$	$\mathbf{w}^k_{IEM}$	$\mathbf{w}^k_{GAM}$
$\mathbf{w}^k$	3,2675	0,4200	1,8600
$\mathbf{w}^k_{AEM}$		2,7250	3,9475
$\mathbf{w}^k_{IEM}$			1,6125

Table 6 Average SRC based on 100 7×1 priority rank vectors

Average SRC (7×1)			
	$\mathbf{w}^k_{AEM}$	$\mathbf{w}^k_{IEM}$	$\mathbf{w}^k_{GAM}$
$\mathbf{w}^k$	0,7093	0,9650	0,8393
$\mathbf{w}^k_{AEM}$		0,7425	0,6746
$\mathbf{w}^k_{IEM}$			0,8668

Table 7 Average SRC based on 100 8×1 priority rank vectors

Average SRC (8×1)			
	$\mathbf{w}^k_{AEM}$	$\mathbf{w}^k_{IEM}$	$\mathbf{w}^k_{GAM}$
$\mathbf{w}^k$	0,6888	0,9600	0,8229
$\mathbf{w}^k_{AEM}$		0,7405	0,6240
$\mathbf{w}^k_{IEM}$			0,8464

greatest mean differences were found between vectors corresponding to the inconsistent matrices and matrices adjusted with the AEM. Tables 6 and 7 display the Spearman correlation coefficients for ranks, to point out the correlation between the derived priority weight rank vectors of the inconsistent starting matrices and their adjustments. The results demonstrate, that the highest positive correlations can be observed after adjustment via the IEM and the GAM.

In short, the main findings can be outlined as follows:

With respect to priority weight rank vectors the IEM produces the most similar vectors after adjustment. Here, the GAM gets only the second place.

But using the GAM seems to be a reasonable alternative to other automated consistency adjustment approaches. The results of the comparisons show, that the matrices adjusted with the GAM are closest to the underlying matrices of the decision makers in terms of MSME. Under the assumption, that a decision maker has made his evaluations with care, one would expect, that he would prefer an adjusted solution which incorporates his evaluations as good as possible.

Furthermore the adjusted matrices calculated by using the GAM contain only entries the values of which belong to the Saaty scale, i.e. they correspond to the accompanying verbal expressions which would make it easier for a decision maker to decide, if he would agree with a adjusted solution.

4 Outlook

Research on consistency adjustments for comparison matrices is a very important field of AHP related research. This study of different approaches gives first insights, which methods to which extent might be appropriate for automated consistency adjustments. Further research could be conducted with respect to other levels of inconsistencies. Furthermore, the GAM opens a door to conduct an empirical study, where decision makers are confrontated with adjusted matrices and can decide to which extent they would accept the adjusted comparisons.

References

Carmone, F. J., Kara, A., & Zanakisc, S. H. (1997). A Monte Carlo investigation of incomplete pairwise comparison matrices in AHP. *European Journal of Operational Research, 102*(3), 538–553.

Decker, R., Meissner, M., & Scholz, S. W. (2008). Detecting and debugging erroneous statements in pairwise comparison matrices. In: J. Kalcsics, and S. Nickel (Eds.), *Operations Research Proceedings 2007* (pp. 277–282). Heidelberg: Springer.

Harker, P. T. (1987). Alternative modes of questioning in the analytic hierarchy process. *Mathematical Modelling, 9*(3–5), 353–360.

Ho, W. (2008). Integrated analytic hierarchy process and its applications: A literature review. *European Journal of Operational Research, 186*(1), 211–228.

Ishizaka, A., & Lusti, M. (2004). An expert module to improve the consistency of AHP matrices. *International Transactions in Operational Research, 11*, 97–105.

Lin, C.-C., Wang, W.-C., & Yu, W.-D. (2008). Improving AHP for construction with an adaptive AHP approach. *Automatation in Construction, 17*(2), 180–187.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *The Psychological Review, 63*, 81–97.

Saaty, T. L. (1980). *The analytic hierarchy process*. New York: McGraw-Hill.

Saaty, T. L. (2003). Decision-making with the AHP: Why is the principal eigenvector necessary? *European Journal of Operational Research, 145*(1), 85–91.

Saaty, T. L., & Ozdemir, M. S. (2003). Why the magic number seven plus or minus two. *Mathematical and Computer Modelling, 38*(3–4), 233–244.

Vaidya, O. S., & Kumar, S. (2006). Analytic hierarchy process: An overview of applications. *European Journal of Operational Research, 169*(1), 1–29.

Vargas, L. (1982). Reciprocal matrices with random coefficients. *Mathematical Modelling, 3*(1), 69–81.

Zanakis, S. H., Solomon, A., Wishart, N., & Dublish, S. (1998). Multi-attribute decision making: A simulation comparison of select methods. *European Journal of Operational Research, 107*(3), 507–529.

Zeshui, X., & Cuiping, W. (1999). A consistency improving method in the analytic hierarchy process. *European Journal of Operational Research, 116*(2), 443–449.

Monte Carlo Methods in the Assessment of New Products: A Comparison of Different Approaches

Said Esber and Daniel Baier

Abstract In New Product Development (NPD), different financial assessment methods can be used, e.g., static and dynamic net present value methods, decision tree methods, and real options approaches. In all approaches, uncertainties can be specified and their impact on the project value can be analyzed using Monte Carlo Methods. This paper describes the use of Monte Carlo methods in information technology (IT). For comparing these different methods, the NPD case of the new video-conference system BRAVIS is elaborated.

1 Introduction

There are strategic reasons for new product developments. Successful, new product developments can offer long-term, financial returns on investments which support marketing efficiency, increase corporate image and use production, operative as well as human resources effectively. Investments in product development are of prime importance to research & development (R&D) projects. On the one hand, these investments hold a strategic position because they shall ensure and increase the future competitiveness of the company. On the other hand, these investments highly influence the financial strength of the company because of being linked with high expenditures. The financial assessment of product development projects includes the uncertainties and acting possibilities of the management. The characteristics of R&D projects cannot be considered certain because these projects are future-related. From a corporate point of view, it is distinguished between technical risks (realisation risk, performance risk, cost risk and regulatory risks) and economic risks (price risk, competition risks, Me-Too-products and the risk of accepting the new product) (Dilling 2002; Pritsch 2000).

S. Esber (✉)
Chair of Marketing and Innovation Management, Brandenburg University of Technology Cottbus, Postbox 101344, 03013 Cottbus, Germany
e-mail: esbersai@tu-cottbus.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_77,

Three possible courses of action are just the right thing for the R&D management in order to react to the uncertainty problem. The first possibility is the conscious or unconscious ignorance of uncertainty aimed at making the decision easier if the information is withheld from the field of management. The second form of reaction is the reduction of uncertainty by gathering information so that the information basis is improved. The third possibility is the acceptance of uncertainty. The analysis of risk structure and flexibility of decisions can make future chances feasible and occurring risks reducible (Adam 1996; Damodaran 2001; Kirchler 1999). Within the scope of this paper, the following research issue is to be addressed: In which way can the practical use of decision-supporting tools (decision trees, Real Option approach and Monte Carlo Simulation) be realized in the assessment of new products (here: Video-conference system) in R&D projects?

2 Assessment Methods of NPD

In R&D projects, various financial methods for assessing new product development are used. The net present value (NPV) method is a standard capital market-oriented method. This method is that monetary procedure of dynamic investment calculation most frequently used. The NPV method considers all monetary movement (in and out) involved in investment and measures the advantageousness of an investment absolutely in a value (net present value). In this respect, net present value is the sum of all monetary movement (in and out) caused by an investment object discounted at any point in time. Within the framework of strategic decisions (takeover and fusion) of companies or product development), massive simplifications of the models and inaccuracies of the results are accepted with the net present value method (Copeland and Antikarov 2001).

Sensitivity analysis can be used for identifying the main risks as far as success of a project is concerned. For this purpose, an optimistic and a pessimistic case are recalculated for each uncertain value driver to jointly fix the interval in which the values can range. Afterwards, the effects of the pessimistic or optimistic assumption are individually calculated for each value driver. A problem in using the sensitivity analysis is that no standardisation for the definitions of “optimistic” and “pessimistic” exist (Brealy and Myers 2003; Damodaran 2001; Franke and Hax 1999).

Monte Carlo Simulation is a tool for risk analysis and depiction of all possible combinations of variables. At first, a model for the calculation of the project value is formulated and the value drivers are identified. After this, a probability distribution is determined by empirical assessments for each of these value drivers. In each simulation sequence, a respective Software-Program (here: @Risk) draws a value for each of the uncertain value drivers using an accident generator and calculates the project value for this combination. While the advantages of this method are the high freedom of data assessment and the modelling of complex problems, the problems of this method will occur if the number of dependencies increases and the

probability distributions for the value drivers cannot be determined (Hommel and Pritsch 1999; Perridon and Steiner 1999).

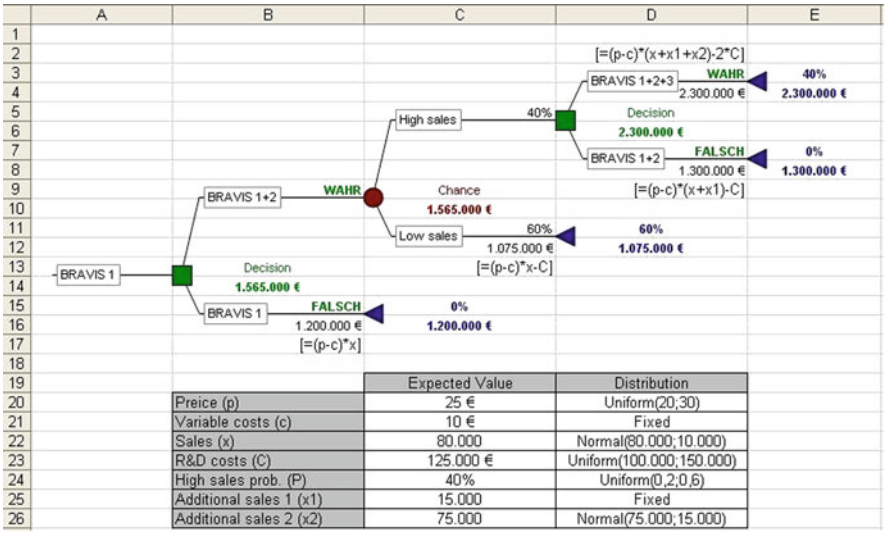
Consideration of the acting flexibility of the management (e.g. additional marketing activities, capacity extensions or sale of the production plant) can be made by the decision tree method. The decision problems can be mapped by decision trees. In this respect, the project sequence is subdivided into various phases. The project is subdivided into various phases considering the fact that during project execution new information becomes available to change the cash value. Whereas decision nodes show the acting possibilities of the management, the state nodes represent the results of acting alternatives in dependence on the occurrence of various surrounding conditions. The assessment of the tree and the determination of the optimum behavior pattern are carried out recursively from right to left. The best decision to be taken is determined at each decision node. The decision trees make thinking about future strategy necessary and represent the connection between the present situation and possible future developments. For this reason, the complexity of decision trees accompanied by a great number of decision points and alternatives are considered a restriction (Brealy and Myers 2003).

The term "real options" was substantiated by Myers in 1977 for the first time (Dixit and Pindyck 1995). A real option means future scopes of activity and investment possibilities of a company, connected with the ability of the management to adjust operative decisions to changed environmental conditions (Hommel and Pritsch 1999). The real options are characterised by five properties: uncertainties of future returns, irreversibility of investment costs, flexibility with regard to determination of execution time of option, exclusivity of exercise right of option only by option owner and information access with regard to improvement of information basis (Brach 2003). In R&D projects, option types are often classified into three groups consisting of learning, insurance and growth options. The learning options make it possible for the company to acquire new knowledge, learn from it and then take the investment decisions on a more solid foundation. The learning options include postponement options (waiting options), delay options and stage investment options (Amram and Kulatilaka 1999; Copeland and Antikarov 2001). The insurance options enable the company to react to unfavourable market developments and thus avoid future losses. For this reason, these options serve for risk management of a company. The insurance options include capacity change options, breaking-off options and readjustment options (Mostowfi 2000; Trigeorgis 1996). The growth options will offer management the possibility of expanding the primary indication if unfavourable environmental conditions occur and keeping as well as improving the competitive position. The growth options include the extension options and innovation options (Brealy and Myers 2003; Kilka 1995). For the assessment of whether an investment makes sense or not, the assessment method should consider some criteria which distinguish a good decision rule from all others, such as guarantee of maximisation of the company value, presentation of the uncertainties and flexibility, consideration of the irreversibility in terms of investments that have already been realised and applicability to various projects and investment projects, respectively.

3 Real Options in the Assessment of NPD

Video-conference means a general form of communication in which people talk to and see each other at the same time even though they are not sitting in the same room. The video-conference system BRAVIS (Brandenburg Video Conferential Service) has been developed at the Chair for Computer Networks and Communication Systems at the Brandenburg University of Technology (BUT) Cottbus and within the framework of a federal state project (Zühlke 2004). BRAVIS shall support telegraphing applications (e.g. teleprinters, telecommunications, examinations and special lectures) and closed user groups of up to 20 participants. For marketing the BRAVIS-approach, the BRAVIS GmbH [www.bravis.eu] as BUT-Spin-off was founded in autumn 2005. Presently, the company employs 15 staff members at the Cottbus site. More than 20 articles at international conferences, three dissertations, three patents and more than 15 graduation papers have been dealing with BRAVIS.

Excel is used as a decision instrument in the economic and financial field. The decision-supporting Excel-based tools offer R&D management a better understanding of the problem structure and a deep background of each decision. For this reason, these tools are very appropriate and useful for R&D management. Two Excel add-one (Precision Tree and @Risk 4.5 from Palisade in Theca, NY) can be used to depict the decision trees, to analyse economic and technical uncertainties and to make sensitivity results visible (Rese and Baier 2007). Precision Tree offers the necessary tools for depicting and analyzing the decision trees. Figure 1 shows a decision tree which is drawn up by Precision Tree within Excel. R&D management has to decide on product development of BRAVIS 2 for the BRAVIS 1 which



	Expected Value	Distribution
Preice (p)	25 €	Uniform(20;30)
Variable costs (c)	10 €	Fixed
Sales (x)	80.000	Normal(80.000;10.000)
R&D costs (C)	125.000 €	Uniform(100.000;150.000)
High sales prob. (P)	40%	Uniform(0,2;0,6)
Additional sales 1 (x1)	15.000	Fixed
Additional sales 2 (x2)	75.000	Normal(75.000;15.000)

Fig. 1 Decision tree for introducing BRAVIS 2 with assumed uncertainties

has been introduced already. Depending on market acceptance of product development (high or low turnover), the management can take a decision on further product developments (BRAVIS 3). The decision on development and market introduction of BRAVIS depends on: BRAVIS-price p, variable costs c, turnover of BRAVIS 1,2 and 3 (for BRAVIS 1 the turnover is assumed as x "low turnover" and, in the introduction of BRAVIS 2 and 3, possible additional turnover for BRAVIS 2 are modeled as x1 and for BRAVIS 3 as x2), R&D costs for BRAVIS-development C and on the probability P of realizing higher turnover after BRAVIS 2 UNC 3 have been introduced. The values are alienated in this section for confidentiality reasons. Figure 1 shows the calculation of the profits for each terminal node of the decision tree. As far as a decision in favor of the introduction of BRAVIS 2 is concerned, some estimates can be made in such a way. For example, the probability of realizing higher sales is assumed to be equally distributed between 20% and 40%. The possible additional sales of BRAVIS 3 are assumed to be distributed on average of 75,000 systems and a standard deviation of 15,000. Using this information, @Risk carries out a Monte Carlo Simulation for the decision tree, analyses each result and graphically illustrates those uncertainties which decision makers compare (see Sect. 4).

4 Monte Carlo Simulation in Assessment of NPD

Mobility and cooperation play a significant part in the design of internet applications. Today the number of mobile phone users amounting to approx. 3 billion clearly exceeds the number of internet users covering about 1 billion. In the course of the last 15 years, mobile communication market has developed to a growth market whose potential is certainly not yet exhausted. Although great efforts in research and development have been made in the last 10 years to develop attractive mobile phone applications, the use of mobile phone services in Europe is mainly concentrated on language communication and SUMS-service. The design of mobile collaborative services and applications makes a smooth transition between different types of networks, integration of existing web services and endurance of relevant protection targets (e.g. confidentiality, privacy) necessary. It must become possible to actively participate in relevant decision processes world-wide and to receive the required information irrespective of whether in a fixed or a mobile network.

Within the framework of the Undereaten Region (Company Region), the innovation initiative for the New Federal States, the Federal Ministry for Education and Research (Undenominational füa Building UNC Forsaking) with its new promotion program Format (Research for the Market in a Team) is now aiming at connecting two new approaches for knowledge and technology transfer. The strategic aim of the Format-project is the development of a platform for providing mobile collaborative applications and services. The characteristic features of this platform are: Openness, because the platform shall be flexible and extendable and offer the possibility of integrating existing web services. The second feature is the transparent use

of the web because the mobile collaborative applications shall be independent of the network and usable by the fixed and mobile networks. The third feature is the safe communication because the platform shall support confidential communication, if required. The three business fields of the project are Mobile Video-based Collaboration (based on BRAVIS), Smart Home on the Phone Applications and Mobile Collaborative Games. Within the framework of this paper, the business field Mobile Video-based Collaboration is assessed. Web or video conferential is considered a helpful means for saving time and costs. They are regarded as a simple instrument for the transfer of information and for efficient communication. At the moment, the international market for audio- and teleconference systems is growing by about 17% per year. Because of the increasing mobility of the communication partners and the fast moving character of the world of work, mobile video communication solutions make it possible that mobile partners participate in video and audio conferences while on the way and can access important documents as well as ensure a stable, qualitatively high connection between the participating partners so that a continuous exchange of information is enabled. The customer groups for this field of business are private customers, clients and founders of a new business for software developments. At the beginning, the product and service offer is directed to business customers who are in need of collaborative solutions. In the further course, the private customer market shall also be opened up. While the sale of licenses is prioritized in the field of business customers, experience gained by BRAVIS GmbH has revealed that high income can be reached in the field of private customers mainly in connection with advertisement pop-ups.

The proportional costs for the development of the service for mobile video-based collaboration amount to 250,000 Euro. The platform to be developed should be provided as a product within 2 years term of the project. In the private field, a charge of 1 Euro for a one-time use of the service, monthly 1,500 users of the service (yearly 18,000 Euro) will accrue. In the business field, 150,000 Euro would already have accrued as cash flow based on a sales price of 150 Euro and a sales volume of 1,000 licenses a year. Further sales can be generated by the founders for software development. For average prices of 10,000 Euro, there will be 13 software developments for further 130,000 Euro a year. A model for calculating the project value has been formulated using an Excel-calculation table. The values are alienated for confidentiality reasons.

Figure 2 shows the result of the Monte Carlo Simulation for the cash flows of the project. In each course of simulation, @Risk draws a value for each uncertain value driver and afterwards the decision tool reflects the possibility of introducing the new product. An @Risk simulation calculates multiple scenarios of a model by repeatedly sampling values from the probability distributions for the uncertain variables and using those values for the cell. @Risk simulations can consist of as many trials (or scenarios), hundreds or even thousands in just a few seconds. During a single trial, @Risk randomly selects a value from the defined possibilities (the range and shape of the distribution) for each uncertain variable and then recalculates the spreadsheet. @Risk offers a sensitivity analysis for determining critical factors. This analysis can be used for the classification of uncertain parameters in dependence on

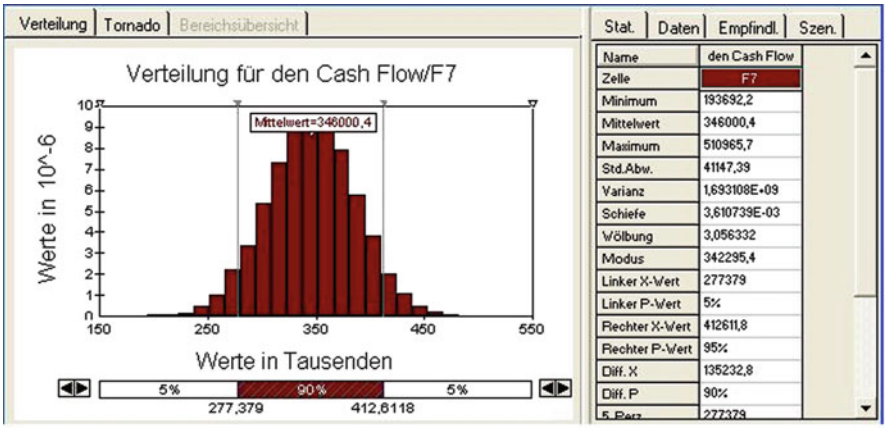


Fig. 2 Distribution of the Cash Flow according to a Monte Carlo simulation

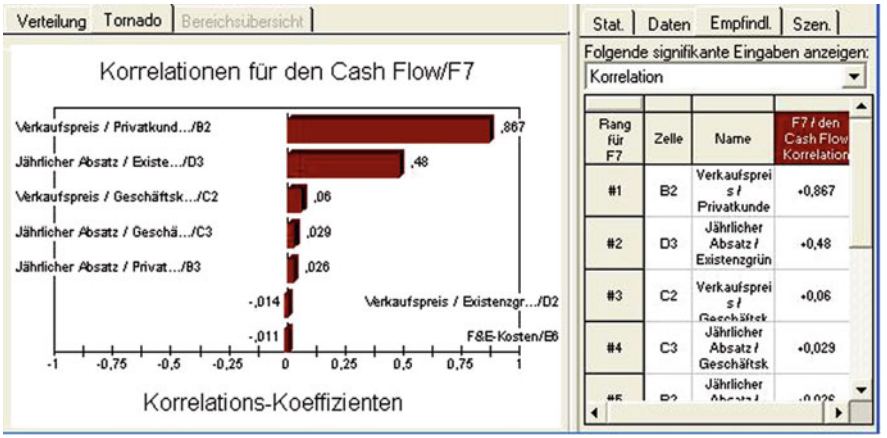


Fig. 3 Sensitivity analysis for the Cash Flow

their effects on profits. Figure 3 shows the sensitivity analysis in form of a tornado diagram. The tornado-diagram shows that the charge for a one-time use of the service in the field of private customers and the sales volume in the field of founders for software development have a very strong influence.

5 Conclusions and Outlook

The Paper shows that the use of the Real Options approach in R&D projects using decision tools is possible and reasonable. The consideration of the Real Options method is very significant for modelling the technical uncertainties, and those

which have been caused by the market during the development processes of the BRAVIS-system, as well as for modelling the possible courses of action done by the management using decision trees and risk analysis. Within the framework of the Monte Carlo Simulation, different distributions for uncertain parameters could be considered. From this, a distribution for the project value could be derived and the most significant value drivers recognisee.

References

Adam, D. (1996). *Planung und Entscheidung*. Wiesbaden: Gabler.

Amram, M., & Kulatilaka, N. (1999). *Real options: Managing strategic investment in an uncertain world*. Boston: Harvard Business School Press.

Brach, M. (2003). *Real options in practice*. Hoboken: Wiley.

Brealy, R. A., & Myers, S. C. (2003). *Principles of corporate finance*. Boston: McGraw-Hill.

Copeland, T., & Antikarov, V. (2001). *Real options: A practitioner's guide*. New York: Texere.

Damodaran, A. (2001). *Corporate finance*. New York: Wiley.

Dilling, A. A. (2002). *Anwendung und Anwendbarkeit der Realoptionstheorie zur Bewertung von Forschungs- und Entwicklungsprojekten unter besonderer Be-rücksichtigung projektendogener Risiken*. Diss. Univ. Göttingen.

Dixit, A. K., & Pindyck, R. S. (1995). The options approach to capital investment. *Harvard Business Review*, 105–115.

Franke, G., & Hax, H. (1999). *Finanzwirtschaft und Kapitalmarkt*. Heidelberg: Springer.

Hommel, U., & Pritsch, G. (1999). Marktorientierte Investitionsbewertung mit dem Realoption-sansatz. *Finanzmarkt- und Portfoliomanagement, 13*(2), 121–144.

Kilka, M. (1995). *Realoptionen: Optionstheoretische Ansätze bei Investitionsentscheidungen unter Unsicherheit.* Frankfurt a. M.: Fritz Knapp Verlag.

Kirchler, E. M. (1999). *Wirtschaftspsychologie*. Göttingen: Hogrefe.

Mostowfi, M. (2000). *Bewertung von Unternehmensakquisitionen unter Berücksichtigung von Realoptionen.* Frankfurt a. M.: Europäischer Verlag der Wissenschaft.

Perridon, L., & Steiner, M. (1999). *Finanzwirtschaft der Unternehmung*. München: Vahlen.

Pritsch, G. (2000). *Realoptionen als Controlling-Instrument*. Wiesbaden.

Rese, A., & Baier, D. (2007). Deciding on new products using a computer-assisted real options approach. *International Journal of Technology Intelligence and Planning, 3*(3), 292–303.

Trigeorgis, L. (1996). *Real options: Managerial flexibility and strategy in re-source allocation.* Cambridge, MA: MIT Press.

Zühlke, M. (2004). *Verteilt organisierte Mehrteilnehmer-Videokonferenzen für geschlossene Gruppen im Internet.* Diss. BTU Cottbus.

Preference Analysis and Product Design in Markets for Elderly People: A Comparison of Methods and Approaches

Samah Abu-Assab, Daniel Baier, and Mirko Kühne

Abstract The elderly people fraction is rapidly increasing in the industrialized countries. The average life expectancy for many Europeans is now over 80, and by 2,020 around 25% of the population will be over 65. It is known that with ageing the behavior changes and body condition starts to diminish. Hence, the urge for new products and services that meet the target group's needs, requirements, inspirations, and comfort is essential and promising for potential producers. As a result, it is a critical issue for market researchers to adjust their methods and techniques to those markets. The main focus of the paper is to suggest a new approach for measuring the preference of elders. The new approach can be used for designing products and services. It is based on well-known preference analysis methods (i.e., a combination of conjoint analysis and quality function deployment as with Baier 1998 or Pullman et al. 2002). The various approaches are discussed and implemented on elders. The development of a new mobile phone for elders is used as a demonstration and validation sample.

1 Introduction

This paper investigates two approaches which combine conjoint analysis (CA) in quality function deployment (QFD) known as conjointQFD.

Quality Function Deployment QFD is a method to develop a design of quality which meets the needs and demands of customers (Akao 1990). Two objectives linked the need of QFD which started with the user (or customer) and the other ended with its producer. The first objective is the conversion of customers demands and requirements – product attributes (PAs) – into product characteristics (PCs). The second objective is the deployment of quality characteristics identified to production

S. Abu-Assab (✉)
Chair of Marketing and Innovation Management, Brandenburg University of Technology Cottbus, Postbox 101344, 03013 Cottbus, Germany
e-mail: samah.assab@tu-cottbus.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_78,

activities (Revelle et al. 1998). During its diffusion three different approaches in the work process was introduced: the comprehensive QFD concept (Akao 1990), with his thirty-table concept besides the four-phase concept from ASI (ASI 1992). The "House of Quality" (HoQ) is the most implemented phase in QFD. In the HoQ, customer demands are collected then they are interpreted to engineering characteristics by an expert team. The subjectivity (i.e., potential error) of collecting the customers voice and determining the relationship of PCs on PAs led the researcher to try to find ways to minimize and to eliminate the source of error (e.g., Hauser and Simmie 1981). Other researchers suggested the use of conjoint analysis (Gustafsson 1996; Baier 1998; Pullman et al. 2002).

Conjoint Analysis Since its introduction into marketing in the early 1970s, CA was considered as a major set of techniques for measuring the preferences of the customers among multi-attributed products and services (see Green and Rao 1971; Srinivasan and Shocker 1973). CA has also gone through a lot of improvements; new models were developed. For example, choice-base conjoint, hybrid conjoint including Johnson's adaptive CA model (Green 1984). In estimating the part-worth, the hierarchical Bayes proved to be preciser than other used methods e.g., the ordinary least square (Green et al. 2001). In the last two decades, researcher have tried to combine and compare other methods with/to CA to improve their results. Baier (1998) used conjointQFD. Baier's result identified that conjointQFD outperforms the traditional QFD. ConjointQFD is designated in the paper as Baier's Approach. Pullman et al. (2002) compared QFD and CA by applying each to the design of a new product: all-purpose climbing harness. Their interpretation of their results showed that the two methods should be considered as complementary than competitive and are better conducted simultaneously. Pullman's experiment is designated in this paper as Pullman's Approach. The paper suggests a new approach tailored for elders and compares its results to Pullman's approach results on the example of mobile phones for elders. The next section will handle the elderly group with its various peculiarities and specifications.

Elderly People The industrial countries are facing a dramatic demographical development in their societies: their societies are growing old. The main aspects of the problem are observed in the increase in the mean and median age, and in the transition in the children fraction in a population, which are also the main causes of the demographical problem (Gavrilov and Heuveline 2003). Growing old or ageing in simple words is "the body's natural process of wearing down." It is an individualized process that begins at birth, and is affected by many factors as people age. In modern gerontology, ageing is defined as a process that takes place at one time at different levels: physically, emotionally, socially and socio-culturally. On one or more of those levels, it can be triggered in which way the process of ageing is initiated and its effect on the other levels (Baltes and Baltes 1994). The process of ageing involvespartly conditional and possible physiological changes that happen to elderly. These changes differentiate the elderly group from other younger groups e.g., N.A. (2009). Despite thesenatural changes, today's elderly generation are not to be compared

with elders from previous generations. Today's elders seek a quality life and self-fulfilment. For the marketers, they build a new confident consumption generation. For example, elders (taken here ab 60plus) expenditure reaches 316 billion Euros, one-third of total expenditure from the household in Germany. According to the Ministry for Family, Seniors, Women and Youth, this share will increase to 386 billion Euros, more than 41% of total household expenditure until 2,050 from purely demographical factor. Product developers and market researchers have to take the elders into consideration when designing and marketing products. The products and innovations should better meet needs of the elderly group to contribute into an easier life for them as long as possible.

2 New Approach for Elderly People

Baier's approach implements CA in HoQ so that to determine the importance of PAs and also to evaluate PCs effect on the PAs, taking advantage of the "trade-off" feature of the method, increasing the objectivity of the evaluation (Baier 1998). The new approach uses the same core methodological idea as Baier's approach: CA is used in QFD taking into account the difficulties of elderly people which vary in their intensity from one person to another (e.g., distorted colors, difficulties in reading the font-size, computer competence, knowledge of theme, etc.). These difficulties of elderly are mainly overcomed in the new approach by introducing an optional observation phase, an obligatory information phase and adjusting the interviews questions in the online questionnaire to simplify the task for the elderly people. For example by using technologies (e.g., animated multimedia, avatars, pop-ups info, videos, voice questions, etc.) to make it easier for them as well as to get more valid results. Table 1 illustrates the two approaches, their main steps and the new integrated levels. From Table 1, the first integrated level in step one is the Information phase which is obligatory and it should be conducted at the very beginning. It supplies the elders with the necessary update of the product options and state-of-the-art of the tested product or service. A slight difference can be recorded in the way of collecting the PAs in both methods (see Table 1). In the new approach, it is recommended to use face-to-face interviews with elderly (e.g., less than 30 respondents. In step two, another information phase is introduced to help those who need more information on the subject. In step 3, the elders group should be presented when building the expert team. In the next step (step 4) an information phase could be implemented in the various CA to calculate the correlation between the PAs and PCs. Methodologically, the new approach doesn't differ from the baier's approach and it should be applicable for elders as well as for all age-groups. The levels added aims to adjust certain defficiencies due to ageing to make it simpler and clearer for elders. The experiment was conducted using the Baier's approach with minimum information phases, therefore for the purpose of the current research the Baier's approach is taken as the new approach. The details of the experiment is discussed in the next section.

Table 1 Baier's approach vs New approach for elderly People

Baier's approach	New approach for elderly
Step1: Selecting PAs	
PAs are selected by QFD team, using info from retailers Journals, internet.	Recommended face-to-face Interviews with integrated observation phase, integrated information phase and optimal presentation of attributes and levels
Step 2: Evaluating the importance of PAs	
Running conjoint analysis	Running conjoint analysis with integrated information phase
Step 3: Selecting PCs	
Expert team	Expert team (includes elderly members)
Step 4: Estimating the influence of PCs on PAs	
Expert team, conjoint analysis	Expert team, conjoint analysis with integrated information phase
Step 5: Computing importance of PCs in the eye of customers	
Sum of PC shares per PA shares weighted by PA shares	Sum of PC shares per PA shares weighted by PA shares

2.1 Application of the New Approach for Elderly People

In a previous work, the authors investigated the Pullman's approach on the example of mobile phone design for elders (see Abu Assab and Baier 2010). In this paper, the new approach is applied on the same example. ECs and PAs were taken from the HoQ from Pullman's approach to guarantee the same conditions in the two experiments (for the comparison purpose). Elders were asked to evaluate both the PAs and their importance as well as the influence of PCs on the PAs. A total of 10 ACA interviews were assessed by each of 40 respondents with average time of 2:15 h/interview. ACA1 was performed to assess the PAs importance. From ACA2-ACA9, the strength of the correlation was assessed between the PAs and PCs. Finally, ACA10 was conducted to test the validity of the results. A Summary of results of the ACAs in the HoQ is shown in Table 2. A disadvantage of the approach is that it requires a long time and elders are then overburden from running the interviews. The results show that elders' most important product characteristics are: "emergency number" (0.111), "weight" (0.095), "cost" (0.087);"Volume" (0.059); "display size" (0.035), whereas; the PCs with the least importance for elders are: "battery capacity" (0.011);"waterproof" (0.021); "sound quality" (0.022); and "distance between keys" (0.024).

Table 2 The result of HoQ with the importance (standard deviation) of mobile phones using the new approach

			Product characteristics																								
		Importance of PAs	↓	↑	↑	↑	↑	↓	○	↑	↑	↑	↑	○	○	○	○	↑	↑	↓	↑	↓	○	↑	↑	↓
			PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20	PC21	PC22	PC23	PC24
			Menu layers	Keys size	Distance between keys	Display size	Display's font size	Reflection of display	Display brightness	sending strength	Receive strength	Sound quality	Sound intensity	SMS/MMS	Reminder function	Help function	Emergency Button	Impact strength	Water proof	Power consumption	Capacity of Battery	Weight	Volume	Color contrast	Keyboard's font size	Cost
Product Attributes (PAs)	PA1	Ease of use	0.122 (0.05)	0.165 (0.07)	0.109 (0.06)	0.072 (0.05)								0.146 (0.09)	0.103 (0.1)	0.134 (0.07)	0.162 (0.09)						0.109 (0.06)			
	PA2	Ease to see	0.127 (0.04)			0.217 (0.12)	0.315 (0.13)	0.230 (0.14)	0.238 (0.13)																	
	PA3	Quality of calling	0.112 (0.04)							0.241 (0.1)	0.299 (0.12)	0.200 (0.13)	0.259 (0.11)													
	PA4	Additional menu	0.071 (0.05)											0.272 (0.16)	0.194 (0.16)	0.200 (0.13)				0.333 (0.13)						
	PA5	Emergency menu	0.091 (0.05)														1.00 (0)									
	PA6	Robustness	0.077 (0.04)															0.337 (0.18)	0.278 (0.2)			0.385 (0.2)				
	PA7	Long Battery	0.080 (0.04)	0.150 (0.06)		0.089 (0.05)			0.072 (0.05)	0.081 (0.05)	0.080 (0.04)		0.099 (0.05)	0.117 (0.07)	0.078 (0.05)					0.091 (0.04)	0.144 (0.05)					
	PA8	Mobile's Manageability	0.112 (0.05)																			0.586 (0.17)	0.414 (0.17)			
	PA9	Easy to read Keyboard	0.122 (0.04)		0.290 (0.11)	0.123 (0.11)																		0.273 (0.15)	0.314 (0.12)	
	PA10	Price	0.087 (0.05)																							1.00 (0)
	Importance of PCs		.032	.049	.024	.035	.040	.029	.036	.034	.040	.022	.037	.047	.033	.031	.111	.026	.021	.031	.011	.095	.059	.033	.038	.087
	Rank		16	5	21	12	7	19	11	13	8	22	10	6	15	18	1	20	23	17	24	2	4	14	9	3

Table 3 Ten most important PAs: Pullman's approach, new approach and ACA

	10 Most Important Product Characteristics (PAs)									
Importance (rank)	Size of keys	Emergency	Impact strength	Water proof	Energy consumption	Battery capacity	Weight	Volume	Cost	
Pullman's approach	0.733	0.653	0.725	0.921	0.921	0.972	1.00	0.608	0.617	0.900
	(5)	(7)	(6)	(3)	(3)	(2)	(1)	(9)	(8)	(4)
New approach	0.032	0.049	0.111	0.026	0.021	0.031	0.011	0.095	0.059	0.087
	(6)	(5)	(1)	(8)	(9)	(7)	(10)	(2)	(4)	(3)
ACA	0.127	0.066	0.125	0.0819	0.069	0.102	0.117	0.099	0.061	0.152
	(2)	(9)	(3)	(7)	(8)	(5)	(4)	(6)	(10)	(1)

2.2 Comparison of Results

The top ten product Characteristics for elderly people are compared by Pullman's approach, new approach and the ACA. The results of Pullman's approach shows that "battery capacity" rank(1) is the most important PC, "energy consumption" rank(2), "water proof" and "impact resistance" both rank(3). On the other hand, the results of the new approach show that the most important PCs are "emergency number" rank(1), "weight" rank(2), and "cost" rank(3). ACA results are as follows "cost" rank(1),"menu layers" rank(2), followed by "emergency button" rank(3) (see Table 3 for the top 10 results of the three approaches). The overall results from the comparison show that there are some similarities, however, the differences are significant. In this experiment, the authors want to check the external validity: Which approach has a better external validity? For this purpose, a "field predictability test" is conducted to the validity. The test consists of various measures that test real life mobiles for the three approaches to figure out which approach outperforms in real life (see next section).

2.3 Field Predictability Test

The expert team selected 20 mobile phones from the several classes for elderly (e.g., "AURO Comfort 1010", "Emporia Life Plus", "Nokia 1200", "Nokia 6500 slide" and "Motorola RAZR2 V8"). The most ten top PCs were selected as the test criteria to perform the "field predictability test" (see Table 3). Then each mobile phone was assigned a value for each selected PA using its product catalogue. These specifications were then evaluated for each mobile using Pullman's approach, new approach and ACA. A case in point is the "Emporia Life Plus": PC values were assigned as follows (average values): "emergency number": yes (P:12.74, B: 58.93, ACA: 14.35) "Key size": big (P: 19.65, B:47.74, ACA: 54.26). This procedure enabled us to rank the 20 mobile for each approach. In this step, three evaluation measures: hits in "Google.de", number of reviews in "amazon.de" and number of offers in

Table 4 Field predictability test total results

	Pullman Approach	New Approach	ACA
Hits in "google.de"	1	3	1
Reviews in "amazon.de"	3	3	2
Offers in "amazon.de"	1	1	1
Totals	5	7	4

"amazon.de" were selected to run the comparison. The new approach recorded more hits than from other approaches: 5, 7, and 4 hits respectively (see Table 4).

3 Discussion and Outlook

The main goal of the paper is to introduce a new approach for elders and then test its external validity in comparison to the other approaches mentioned in the paper. Methodologically, the new approach is similar to the Baier's approach (see Chapter 1) with some adjustment phases to pass with the normal process of ageing. The "field predictability test" was introduced. Accordingly, the external validity was tested for the various approaches using internet indicators. The main result shows that the new approach correlate better as Pullman's approach and the traditional ACA. However, the disadvantage of the new approach – it requires long time from elders- overburden the elderly people. Therefore, further research is recommended to improve the measures -better test the predictive validity- as well as to check how reasonable is the new approach.

References

Abu Assab, S., & Baier, D. (2010). Designing products using quality function deployment and conjoint analysis: A comparison in a market for elderly people, *Studies in Classification, Data Analysis, and Knowledge Organization*, *38*, 515–526.

Akao, Y. (1990). *Quality function deployment: Integrating customer requirements into product design*. Cambridge, MA: Productivity Press.

ASI (1992). Quality function deployment: Executive briefing. Dearborn, MI: ASI Press.

Baier, D. (1998). Conjoint analytische Lösungsansätze zur Parametrisierung des House of Quality In *QFD Produkte und Dienstleistungen marktgerecht gestalten.* Dusseldrof: VDI.

Baltes, P. B., & Baltes, M. M. (1994). Gerontologie: Begriff, Herausforderung und Brennpunkte, in Alter und altern, ein interdisziplinärer Studientext zur Gerontologie (Sonderausgabe des 1992 in 5). Forschungsbericht der Akademie der Wissenschaft zu Berlin, Berlin, 1–34.

Gavrilov, L. A. & Heuveline, P. (2003). Aging of Population. In P. Demeny & G. Mcnicoll (Eds.), The Encylopedia of Population, NY.

Green, P. E. (1984). Hybrid models for conjoint analysis: An expository review. *Journal of Marketing Research*, *21*(2), 155–169.

Green, P. E., & Rao, V. R. (1971). Conjoint measurement for quantifying judgmental data. *Journal of Marketing Research, 8*(3), 355–363.

Green, P. E., Krieger, A. M., & Wind, Y. J. (2001). Thirty years of conjoint analysis: Reflections and prospects. *Interfaces*, 31 (3- Part 2 of 2), 56–S73.

Gustafsson, A. (1996). Customer focused product development by conjoint analysis and QFD. Unpublished Thesis (doctoral), Linköping University, 1996.

Hauser, J. R., & Simmie, P (1981). Profit maximizing perceptual positions: An integrated theory for the selection of product features and price *Management Science*, *27*(1), 33–56.

N. A. (2009). Nursing home lawyer: Protecting the well being of nursing home residents. http://www.nursinghomelawyer.com/nursing_home_law_firm/nursing_home_research/aging_disease/normal_aging_changes.htm>, 01.04.2009.

Pullman, M. E., Moore, W. L., & Wardell, D. G. (2002). A comparison of quality function deployment and conjoint analysis in new product design. *Journal of Product Innovation Management*, *19*(5), 354–364.

Revelle, J. B. M., John W., & Cox, C. A. (1998). The QFD handbook. Toronto: Wiley.

Srinivasan, V., Shocker, Allan D., & Weinstein, A. G. (1973). Measurement of a composite criterion of managerial success. *Organizational Behaviour & Human Performance*, *9*(1), 147–167.

Usefulness of A Priori Information about Customers for Market Research: An Analysis for Personalisation Aspects in Retailing

Michael Brusch and Eva Stüber

Abstract Many studies use a priori information about the customers, either as criteria for the selection of respondents or during data analysis and interpretation. Although the often use of this information, the question is how reasonable it is. We analyse this in the framework of personalisation aspects in the retail market. While we extend the data of a conjoint analysis through Hierarchical Bayes estimation and market segmentation approaches we compare three types of data (a priori data, priority data and benefit data) and their effects with respect to predictive validity.

1 Introduction

Within market research, a lot of studies are based on a priori information about the customers. This information can be used as criteria for the selection of respondents (e.g., "quota sampling") or during data analysis and interpretation (see, e.g., DeSarbo and Mahajan 1984). Although this a priori information is often used, the question is how reasonable it is.

This will be analysed in the framework of personalisation aspects in the retail market. The data of a conjoint analysis (see, e.g., Green et al. 2001) is extended through Hierarchical Bayes (HB) estimation (see, e.g., Lenk et al. 1996) and market segmentation approaches. We compare three types of data: a priori data (information about respondents), benefit data (part worth estimates) and priority data (importance estimates) and their effects on validity. Finally we will identify the best way to estimate preferences in such heterogenous and/or unknown fields as personalisation aspects in retailing.

M. Brusch (✉)
Institute of Business Administration and Economics, Brandenburg University of Technology Cottbus, Postbox 101344, D-03013 Cottbus, Germany
e-mail: m.brusch@tu-cottbus.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_79,

We briefly introduce personalisation aspects (Sect. 2) as well as conjoint analysis, HB and clustering approaches for preference estimation (Sect. 3). Later our empirical investigation (Sect. 4) is shown.

2 Personalisation Aspects in Retailing

The retail market can be differentiated in online shopping and traditional retailing. Although the relatively new alternative of online shopping has a growing importance, its impersonality is still an impediment (see, e.g., Holzwarth et al. 2006). A long time traditional retailing was dominated of impersonal shops (e.g., hypermarkets). Nowadays the opportunity for one-to-one marketing is reopened (e.g., while using debit cards).

Personalisation is defined as using personal advisers, namely welcoming and offering of recommendations to buy (offers of alternatives or additional products) in online shopping or in traditional retailing. Personalisation has a positive impact on customer satisfaction (see, e.g., Mittal and Lassar 1996) and likewise on customer loyalty (see, e.g., Srinivasan et al. 2002).

Personal advisers in online shopping (so-called avatars) lead to more satisfaction with the retailer, a more positive attitude toward the product and a greater purchase intention (see, e.g., Holzwarth et al. 2006). In traditional retailing the customer loyalty is significantly affected by the relationship between customer and seller as well as the quality of this relationship (see, e.g., Dixon et al. 2001).

Recommendations to buy are especially in online shopping of interest. Current investigations focus on the applied technology, the so-called recommender systems and their algorithms (see, e.g., Bodapati 2008). Computer-based recommendations have an impact on the choice of products, customer satisfaction and loyalty (see, e.g., Senecal and Nantel 2004), while time for gaining product information is being decreased and buying decision's quality can be increased (see, e.g., Häubl and Trifts 2000).

3 Preference Estimation in Market Research

To analyse the preferences of consumers (e.g., regarding different personalisation aspects in retailing) conjoint analysis can be used. Conjoint analysis is a method to estimate the structure of consumer's preferences (see, e.g., Green and Srinivasan 1978). Typically, hypothetical concepts for products or services (attribute level combinations) are presented to and rated by a sample of consumers in order to estimate part worths for attribute levels from a consumer's point of view. Nowadays for conjoint analysis a huge number of applications are known as well as many specialized tools for data collection and analysis have been developed (see, e.g., Green et al.

2001). Especially Hierarchical Bayes (HB) estimation and clusterwise estimation procedures seem to be attractive newer developments.

HB estimates individual part worth distributions by "borrowing" information from other individuals (for further descriptions of this aspect in a conjoint analysis setting see, e.g., Baier and Polasek 2003). Preference heterogeneity is not assumed via introducing segments. Instead, the deviation of the individual part worth distributions from a mean part worth distribution is derived from the collected individual data (for methodological details and new developments see, e.g., Lenk et al. 1996; Liechty et al. 2005). The application of HB estimation seems to outperform traditional models w.r.t. to predictive validity and seems to be robust (see, e.g., Orme 2000).

Furthermore, part worth estimation can be enhanced using clustering and clusterwise estimation approaches that provide traditional ways to model preference heterogeneity in conjoint analysis (see, e.g., Baier and Gaul 1999; Brusch and Baier 2008).

We use clustering and clusterwise HB procedures which will be applied to preference data for personalisation aspects. We compare three types of data (a priori, benefit and priority data). This is in contrast to other empirical studies which either don't use conjoint analysis (for measuring preference information, see, e.g., White et al. 2008) or select the respondents from a small variety of demographic data (with a limited number of a priori information, mostly students, see, e.g., Iyengar et al. 2008). We are in search of ways to avoid typical problems in market research like focusing the wrong sample or getting less useful (less valid) results. Especially while investigating personalisation aspects the problem of customers with a heterogenous and/or unknown preference structure (at the individual and the aggregated level) is very prominent.

4 Empirical Investigation

4.1 Research Object and Design

For our investigation most important personalisation attributes are used. In total six attributes (e.g., *type of personalization, place of personalization*) each with three (one attribute) or two (five attributes) levels (e.g., *namely welcoming, personalisation at the internet store*) are integrated. In total, 13 part worth parameters have to be estimated in our analyses.

A conjoint study is carried out using the nowadays standard tool for conjoint data collection, Adaptive Conjoint Analysis (ACA) of Sawtooth Software's ACA system (see, e.g., Sawtooth Software 2002), to be precise ACA/Web as online version of ACA within SSI/Web. For our investigation a four-step analysis is done.

As **step 1** we analyse the overall quality of our study. We had 538 finished questionnaires. Standard ACA methodology is used for individual part worth estimation.

Standard selection criteria (with respect to predictive validity) reduce the number of usable respondents to 318 with passably good R^2-measures (with a mean of 0.784).

In **step 2** we divide our respondents into groups depending on our three types of data. For our cluster based on a priori data (**step 2a**) we select relevant socio-demographic characteristics (age, sex, income, profession) and carry out a cluster analysis (using Euclidean distances and Ward's method). An optimal solution with three clusters can be found. For our cluster based on benefit data (**step 2b**) we use standardised individual part worths (where the attribute level with the lowest (worst) part worth is becoming 0, the best attribute level combination (combination of the best attribute levels of each attribute) is becoming 1). For our following cluster analysis we use again Euclidean distances and Ward's method and find a solution with four clusters. For our cluster based on priority data (**step 2c**) we divide the respondents into groups in accordance to the attribute with the highest relative importance (e.g., the most important attribute (and therefore the largest group) is *influence on personalisation* with 25.5% (81) respondents). Because every attribute has its admirer all six attributes can be integrated as a cluster.

In **step 3** we compute the distributions of individual part worths via aggregated HB as well as via clusterwise HB part worth estimations. For our analysis, the software ACA/HB (see, e.g., Sawtooth Software 2006) is used as the actual most relevant standard tool for conjoint data analysis (with 5,000 iterations as burn in and 10,000 draws to be used for each respondent). Preprocessing in order to segment the available individual data was done via SAS.

In **step 4** we calculate the predictive validity values. This was considered while questioning on the basis of the integration of a specific holdout task. This task included the evaluation of five (additional) personalisation concepts, similar to the "calibration concepts" of an usual ACA questionnaire.

Predictive validity is measured using the Spearman rank-order correlation coefficient and the first-choice-hit-rate. The Spearman rank-order correlation compares the predicted preference values with the corresponding observed ordinal scale response data from the holdout task. The first-choice-hit-rate is the share of respondents where the stimulus with the highest predicted preference value is also the one with the highest observed preference value.

4.2 Results

First we calculated the predictive validity values for traditional (standardised) ACA part worths. When using individual data of the 318 selected respondents a first-choice-hit-rate of 57.40% and a mean Spearman of 0.580 results.

The validity values shown in Table 1 are based on the HB estimation and are given for the total sample and for the three clusterwise estimations (clustered using the a priori data). The clusters are separated after the membership during the estimation (total sample or segment). The description "in total sample" means that the HB utilities of the respondents were computed by "borrowing" information from

Table 1 Validity values for the total sample and for the clusters based on a priori data for HB estimation

	Total Sample (n = 314[4])	**Cluster 1/3** (n = 166[4])		**Cluster 2/3** (n = 101)		**Cluster 3/3** (n = 30)	
		ITS	IS	ITS	IS	ITS	IS
First-Choice-Hit-Rate [%] (using individual draws, n = 10,000)	57.40	55.16	54.75	61.95	62.38	59.55	58.75
Mean Spearman (using individual draws, n = 10,000)	0.580	0.517	0.514	0.666	0.670	0.622	0.625
First-Choice-Hit-Rate [%] (using individual averages)	60.19	57.23	57.23	66.34	65.35	60.00	60.00
Mean Spearman (using individual averages)	0.611	0.540	0.543	0.701	0.698	0.686	0.665
Total Mean Spearman (ITS / IS) (using individual averages)	**0.611**			**0.609 / 0.608**			

[4]...no. of respondent with missing or invalid holdout data that could not be considered; ITS...in total sample; IS...in segment

the total sample (not only from members of the own segment). Thus, the HB estimation happened for all respondents together, but the validity values for the three clusters were calculated later separately. On the other hand, the description "in segment" means that the HB utilities of the respondents were computed by "borrowing" information only from members of the own segment (clusterwise HB estimation).

The validity values in Table 1 are shown for the computations based on the 10,000 draws (10,000 HB utilities) for each respondent and for the computations based on the mean HB utilities (one HB utility as mean of 10,000 draws (iterations)) for each individual. Furthermore, an overall mean value for the Spearman rank-order correlation coefficient over all clusterwise estimations is given. This helps to compare "in total sample"- or "in segment"-estimations.

As you can see in Table 1 the validity values for the total sample as well as for all clusters are higher than the values for the traditional ACA estimation.

In Table 2 the same values are given when the clustering process uses the benefit data. Similar results (higher than ACA estimation, no significant influence of "in total sample"/"in segment"-estimations) can be found.

In Table 3 the values are given when the benefit data is used and the respondents were divided into groups in accordance to their priority (integrated as relative importance). Most results are the same as in other cases (higher than ACA estimation). But in contrast to the others the "in segment"-estimations lead to the highest mean validity value.

When comparing all the results it can be seen that all validity values for the individual averages based HB estimation are higher than for the ACA estimation, regardless which HB estimation basis ("in total sample" or "in segment") is used. In the case of HB estimation using individual draws, a mixed result with respect to

Table 2 Validity values for the clusters based on benefit data for HB estimation

	Cluster 1/4 (n = 110[2])		**Cluster 2/4** (n = 45)		**Cluster 3/4** (n = 93[1])		**Cluster 4/4** (n = 66[1])	
	ITS	IS	ITS	IS	ITS	IS	ITS	IS
First-Choice-Hit-Rate [%] (using individual draws, n = 10,000)	52.65	52.40	53.26	51.76	69.97	69.63	50.40	49.65
Mean Spearman (using individual draws, n = 10,000)	0.547	0.544	0.626	0.623	0.704	0.704	0.428	0.427
First-Choice-Hit-Rate [%] (using individual averages)	54.55	53.64	55.56	48.89	74.19	73.12	53.03	53.03
Mean Spearman (using individual averages)	0.575	0.560	0.664	0.648	0.740	0.725	0.452	0.464
Total Mean Spearman (ITS / IS) (using individual averages)				**0.611 / 0.601**				

[1/2]...no. of respondent with missing or invalid holdout data that could not be considered; ITS...in total sample; IS...in segment

Table 3 Validity values for the clusters based on priority data for HB estimation

	Cluster 1/6 (n = 76)		**Cluster 2/6** (n = 52[1])		**Cluster 3/6** (n = 63[1])	
	ITS	IS	ITS	IS	ITS	IS
First-Choice-Hit-Rate [%] (using individual draws, n = 10,000)	57.32	56.13	56.66	56.10	52.26	52.23
Mean Spearman (using individual draws, n = 10,000)	0.569	0.558	0.573	0.578	0.529	0.537
First-Choice-Hit-Rate [%] (using individual averages)	61.84	64.47	57.69	55.77	52.38	53.97
Mean Spearman (using individual averages)	0.618	0.623	0.609	0.605	0.548	0.555
Total Mean Spearman (ITS / IS) (using individual averages)				**0.611 / 0.616**		

	Cluster 4/6 (n = 20)		**Cluster 5/6** (n = 79[2])		**Cluster 6/6** (n = 24)	
	ITS	IS	ITS	IS	ITS	IS
First-Choice-Hit-Rate [%] (using individual draws, n = 10,000)	66.23	66.77	58.49	58.85	61.77	61.39
Mean Spearman (using individual draws, n = 10,000)	0.595	0.582	0.605	0.619	0.668	0.661
First-Choice-Hit-Rate [%] (using individual averages)	70.00	70.00	63.29	62.03	62.50	62.50
Mean Spearman (using individual averages)	0.621	0.611	0.625	0.645	0.706	0.689
Total Mean Spearman (ITS / IS) (using individual averages)				**0.611 / 0.616**		

[1/2] ... no. of respondent with missing or invalid holdout data that could not be considered; ITS...in total sample; IS...in segment

validity can be found. Furthermore, the usage of priority data leads to the highest validity values.

5 Conclusion and Outlook

We investigated the use of clustering and clusterwise HB as well as combined estimation procedures which has been applied to preference data for personalisation aspects. While we compare the use of three types of data (a priori data, benefit data and priority data) we find hints to avoid some typical problems in market research.

Our study – which is exemplary for investigations in new and unknown market research fields – suggests two implications if personalisation aspects are analysed. Firstly, the usage of HB estimation is recommended. Secondly, the usage of a priori data can be enhanced through the (alternative) usage of priority data (here integrated as the relative importance coming from the estimated preference structure).

Furthermore, for market research (in general) the usefulness of using a priori information (demographic data) can be certified and especially the possibility of focusing priority information must be increased. For conjoint analysis (in special) beside the usage of Hierarchical Bayes estimation the separate HB estimation at the individual cluster level ("in segment") can be confirmed.

References

Baier, D., & Gaul, W. (1999). Optimal product positioning based on paired comparison data. *Journal of Econometrics, 89*(1–2), 365–392.

Baier, D., & Polasek, W. (2003). Market simulation using bayesian procedures in conjoint analysis. In M. Schwaiger & O. Opitz (Eds.), *Exploratory Data Analysis in Empirical Research* (pp. 413–421). Berlin: Springer.

Bodapati, A. V. (2008). Recommendation systems with purchase data. *Journal of Marketing Research, 45*(1), 77–93.

Brusch, M., & Baier, D. (2008). Conjoint analysis for complex services using clusterwise Hierarchical Bayes procedures. *Studies in Classification, Data Analysis and Knowledge Organization, 35*, 431–438.

DeSarbo, W. S., & Mahajan, V. (1984). Constrained classification: The use of a priori information in cluster analysis. *Psychometrika, 49*(2), 187–215.

Dixon, A. L., Spiro, R. L., & Jamil, M. (2001). Successful and unsuccessful sales calls: Measuring salesperson attributions and behavioral intentions. *Journal of Marketing, 65*(3), 64–78.

Green, P. E., & Srinivasan, V. (1978). Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research, 5*, 103–123.

Green, P. E., Krieger, A. M., & Wind, Y. (2001). Thirty years of conjoint analysis: Reflections and prospects. *Interfaces, 31*(3), S56–S73.

Häubl, G., & Trifts, V. (2000). Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing Science, 19*(1), 4–21.

Holzwarth, M., Janiszewski, C., & Neumann, M. M. (2006). The influence of avatars on online consumer shopping behaviour. *Journal of Marketing, 70*(4), 19–36.

Iyengar, R., Jedidi, K., & Kohli, R. (2008). A conjoint approach to multipart pricing. *Journal of Marketing Research, 45*(2), 195–210.

Lenk, P. J., DeSarbo, W. S., Green, P. E., & Young, M. R. (1996). Hierarchical Bayes conjoint analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs. *Marketing Science, 15*(2), 173–191.

Liechty, J. C., Fong, D. K. H., & DeSarbo, W. S. (2005). Dynamic models incorporating individual heterogeneity: Utilityevolution in conjoint analysis. *Marketing Science, 24*, 285–293.

Mittal, B., & Lassar, W. M. (1996). The role of personalization in service encounters. *Journal of Retailing, 72*(1), 95–109.

Orme, B. (2000). Hierarchical Bayes: Why all the attention? *Quirk's Marketing Research Review*.

Sawtooth Software. (2002). ACA system: Adaptive conjoint analysis. Version 5.0. *Technical paper series*, Sawtooth Software.

Sawtooth Software. (2006). The ACA/Hierarchical Bayes v3.0 Technical Paper. *Technical paper series*, Sawtooth Software.

Senecal, S., & Nantel, J. (2004). The influence of online product recommendations on consumers' online choices. *Journal of Retailing, 80*(2), 159–169.

Srinivasan, S. S., Anderson, R., & Ponnavolu, K. (2002). Customer loyalty in E-Commerce: An exploration of its antecedents and consequences. *Journal of Retailing, 78*(1), 41–50.

White, T. B., Zahay, D. L., Thorbjørnsen, H., & Shavitt, S. (2008). Getting too personal: Reactance to highly personalized email solicitations. *Marketing Letters, 19*, 39–50.

Importance of Consumer Preferences on the Diffusion of Complex Products and Systems

Sabine Schmidt and Magdalena Missler-Behr

Abstract Complex products and systems (CoPS) are project-developed, represent a significant proportion of gross value and competitive advantages. A British study investigated the importance of CoPS and found out, that in UK an account for around 21 per cent of gross value added, approximately 133 billion in output, is produced (Davies and Hobday 2005). In UK CoPS count around 15 percent of international trade over past 30 years (Davies and Hobday 2005). CoPS belong to highly innovative products and are specific customised systems. The high product complexity is a result from numerous product components and their interactions. These characteristics lead to other requirements for enterprises and consumers. In the case of CoPS firms have to anticipate individual consumer preferences which influence the success of the diffusion process. Consumer behaviour is the result of various sociodemographic, social, economical and psychological factors (Chandrasekaran and Tellis 2008). In order to investigate the interactions between consumer behaviour and diffusion process of CoPS, it is used the methodology of System Dynamics. A System Dynamics model operates as a decision support system and takes into consideration feedback, time delays and nonlinearities. The model describes the interaction between diffusion process and specific important consumer preferences like income, personal innovativeness and leapfrogging behaviour. An empirical data analysis hereby supports the development of the diffusion model. Different scenarios are used to deepen understanding.

1 Motivation

The development of CoPS is not a new research field. However the development and commercialization of CoPS in the business-to-business-to-consumer sector poses a challenge for enterprises. The reason is the structure and accordingly the features

S. Schmidt (✉)
Chair of Planning and Innovation Management, Brandenburg University of Technology Cottbus, Konrad-Wachsmann-Allee 1, 03046 Cottbus, Germany
e-mail: schmidts@tu-cottbus.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_80,

of such CoPS. Furthermore they have an added value and are very cost-intensive. CoPS belong to capital goods, are customised, have various neteffects and generally are produced in one-off projects. The structure of CoPS, for example their product complexity and the high degree of innovativeness, leads usually to a deceleration of diffusion process (Davies 1997; Tidd et al. 2005). The majority of innovation and marketing literature concentrates on simple products, which are produced in large-scale production. The understanding of consumer behaviour and a detailed analysis of interaction between diffusion of CoPS and consumer behaviour is essential for marketing activities. In the case of cost-intensive and high-technology products it is useful considering socio-economical, behaviour-oriented as well as psychological criteria for instance technical attitude, income and leapfrogging behaviour of consumers. Examples for CoPS are intelligent buildings and business information networks.

Based on the raised research deficiency, the existing investigation considers the feedback between diffusion process of CoPS and consumer behaviour by using System Dynamics modelling.

2 Specifics of Complex Products and Systems

Before accurating the understanding of the major problem and deducing recommendations for action, it is essential to point out the characteristics of CoPS. At present there is no definition of CoPS in general. However it is possible to describe the fundamental characteristics of CoPS. First, CoPS have a systemic structure. That means CoPS consist of numerous elements, sub-systems and are organized in hierarchical structures. Due to this fact system elements by themselves do not achieve their functionality. The second characteristic is derived from the numerous elements. There is a specific interaction between components. The interaction indicates the interface between the components. As a result minimal changes in one component have an important influence on the functionality of the product. In extreme case there are disastrous failures (Hobday et al. 2000; Singh 1997). For a better understanding the speciality of CoPS is presented in comparison with simple products. Simple products have an easy structure, standardized components and are produced in mass production. In general only one enterprise manufactures it. Enterprise does not involve the consumer in the development as well as in the production process. In contrast to simple products it is a matter of urgent necessity that enterprises have to involve the consumer in the development process of CoPS. In conclusion consumer is able to follow up the utility of the CoPS in a better way. The differentiation into supplier-side and consumer-side characteristics is deepening the discrepancy between simple products and CoPS by considering the aspects of product life cycle, use of marketing-mix-instruments, purchase decision, or ex ante and ex post purchase uncertainty et cetera (Schmidt 2009).

3 The Diffusion of Complex Products and Systems

The diffusion research interfaces the product innovation with the consumer behaviour research (Trommsdorff and Steinhoff 2007).The current status of diffusion models has to be analyzed in order to consider the diffusion of CoPS. Due to the fact that CoPS has numerous components with various interactions, it is necessary to select such diffusion models which integrate these characteristics. A total of 93 diffusion models (from 1969 to 2008) were investigated. Within this quantity, diffusion models of products with interations are very rare. In literature such models are named "multi-product growth models", "growth models for multiproduct interactions" or "multi-innovation diffusion models" (Bayus et al. 2000; Mahajan and Peterson 1985; Peterson and Mahajan 1978). Equations by Peterson and Mahajan (1978) are the foundation for describing the diffusion of CoPS because they consider four different interactions separately between two products [see (1)]. Equations are based on principles of the diffusion model by Bass (1969).

$$n_i(t) = \frac{dN_i(t)}{dt} = (\alpha_i + \beta_i N_i(t) + c_i N_j(t))(\bar{N}_i - N_i(t))i \tag{1}$$

The solution of the differential equation $n_i(t) = \frac{dN_i(t)}{dt}$ describes the consumer, who has already bought the product i in a certain period t. $N_i(t)$ is the cumulative amount of adopters who have already bought the product i in a certain period and $\bar{N}$ is the market potential of product i. In this context α is in accordance with coefficient of innovation and β with coefficient of imitation. Equation (1) has the basic structure for describing a CoPS. In order to investigate diffusion process of CoPS it is essential to analyse at least two interactions and two product components. CoPS belong to the value added products. Therefore complementary and contingent interactions are important. **Complementary interaction** means that the coefficient c_i has a positive sign ($c_i > 0$) and in the case of two products, one product has a positive effect on the other one and the other way round. **Contingent interaction** [see (2, 3)] means that there is a strict condition, the adopters of the first product will not buy the second without the first one, for example digital camera and photo viewer.

$$n_1(t) = \frac{dN_1(t)}{dt} = (\alpha_1 + \beta_1 N_1(t))(\bar{N}_1 - N_1(t)) \tag{2}$$

$$n_2(t) = \frac{dN_2(t)}{dt} = (\alpha_2 + \beta_2 N_2(t))(N_1(t) - N_2(t)) \tag{3}$$

These two described positive interactions are the foundation for modelling the diffusion process of CoPS. In this approach, CoPS consist of three product components with complementary and contingent interactions. These product components have their individual and limited life cycle with different duration time.

Every product component gets older gradually or gets broken. After a period of time consumer decides if he will buy a new product component or leapfrog the next technological generation. These aspects are included into the approach for developing the simulation model.

The diffusion process is very dynamic and has a lot of feedback and interrelations between system elements and stimulates the system again. The methodology of System Dynamics is an experimental instrument, considers feedback loops, time delays, nonlinearities and dynamical problems over time. All feedback loops form a differential equation system and each variable is described by one formula (Sandrock 2006). System Dynamics use two analytical elements: causal loop diagrams as well as stock and flow diagrams. The last one describes the simulation model. The structure of the basic diffusion dynamics of product component C with complementary interaction is explained by (4) and Fig. 1 which visualizes the stock and flow structure.

$$\begin{aligned} &\textit{Adoption from complementary neteffect product component C on A} \\ &\textit{[households/year]} = \textit{Coefficient of compl neteffect product component C on A} \\ &\textit{*Potential adopters product comp C} * \textit{Adopters product comp C} \\ &\textit{/(Potential adopters product comp C} + \textit{Adopters product comp C)} \end{aligned} \tag{4}$$

As soon as potential adopter has bought one product component, the depreciation of product component starts and in the model it is calculated by a fixed depreciation time. The life of product component corresponds to the tax code for depreciation in Germany. After depreciation time every adopter decides about a new purchase or leapfrogs a technological generation. In this situation the decision depends on the individual innovativeness of consumer. By using the adopter categories of Mahajan et al. (1990) and Rogers (2003) it is distinguished between innovators and imitators. In the model innovators and imitators percentage of repeat purchases is 16% and 84%.

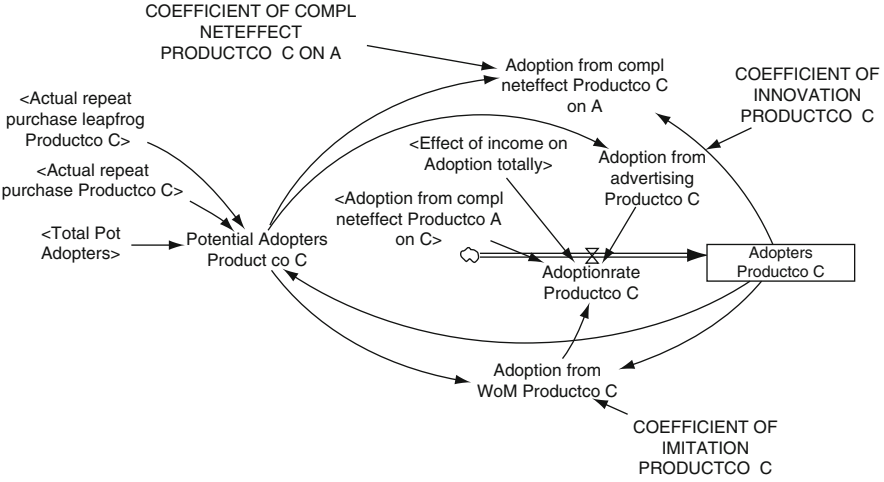


Fig. 1 Basic dynamics for diffusion of product component C

4 Consumer Preferences in the Diffusion of CoPS

This section considers the income of the private households (PHH) in detail. Consumer behaviour is the result of various socio-demographic, psychological, emotional and socio-economical factors (Kroeber-Riel and Weinberg 2003), which control and determine consumer preferences. CoPS are very cost-intensive, the first investment is usually very high and there are other purchases following in order to split expenses. The financial budget of a private household is normally limited, depends on different expenses (Cezanne 2005) and on willingness to pay for a specific product. Experiences of adopters, associations about special product features and income of private household influence willingness to pay. Every private household can allocate its income for expenses after reducing it by tax and contributions by law. The definitions of Statistical Federal Office are the initial point for describing the structure of the model. The income of private households is a stock in a certain period (t) in the System Dynamics methodology. The stock "disposable income of private households" changes by one inflow "the annual income" and one outflow "expenses" [see (6) and Fig. 2].

$$\begin{aligned} \textit{Disposable income PHH [Euro]} = \textit{Disposable income PHH} \\ + \textit{Annual income PHH} - \textit{Expenses PHH} \end{aligned} \tag{5}$$

The annual income is used as a fixed number and is derived from data of the Statistical Federal Office in the period of time 1998 until 2007.

$$\textit{Annual income PHH [Euro/year]} = \textit{Average income PHH} \tag{6}$$

Total expenses consist of private consumption, savings and other expenses (e.g. private health insurance). Besides costs for high-technology products like CoPS have to be added [see (7)]. In dependence of CoPS diffusion process and other influencing factors the expenses for high-technology products increase or decrease. There is a feedback between the diffusion process and the income of private households.

$$\begin{aligned} &\textit{Expenses PHH [Euro/year]} = \textit{Expenses for high technology products} \\ &+ \textit{Consumption} + \textit{Savings} + \textit{Other expenses} \end{aligned} \tag{7}$$

Figure 2 shows the fundamental dynamics for analysing income. For modelling the feedback between income of private households and the diffusion process of CoPS it is used the income elasticity. It gives information about changes of the amount of demand gx in comparison to changes of consumer income gY (Cezanne 2005).

$$\eta(Y) = \frac{gx}{gY} \tag{8}$$

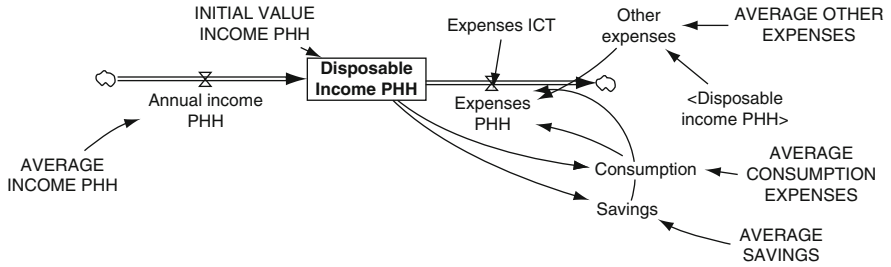


Fig. 2 Basic dynamics for income of private households

5 Model Behaviour

After developing the simulation model it is important to validate its structure and behaviour in order to increase validity and significance (Milling 1974). The existing model is based on scientific approaches and theories in particular diffusion theory of innovations with interactions, consumer behaviour research as well as income elasticity. Simulation model shows the feedback between diffusion process of CoPS, income of private households and individual innovativeness. The simulation model was tested by numerous model structure and model behaviour tests (e.g. structure and parameter verification tests, extreme condition tests, sensitivity tests). Additionally a field research supports configuration of model structure and determination of specific value of model parameters. Out of 1002 private households in Germany (target group) the return rate of the survey was 15.4%.

The validation outcome of all three phases led to satisfying results (validation of model structure, model behaviour and model parameters). The validated model is the basis to generate policies. In order to test strategic decisions there are two possibilities for influencing model behaviour: changes of the parameter values and small changes of the model structure.

Stocks and flows are the relevant analysis variables in System Dynamics models. The simulation period includes 12 years (1998–2009). Numerous product components belong to an intelligent house and it is very difficult to get real data, because of interactions and different life cycles. However three product components belong to the first application field of an intelligent house: personal computer (PC), internet access and digital camera, which were modelled in a contingent and complementary interaction. Real data - the degree of equipment of each product component – were chosen from Federal Statistical Office of Germany.

The private household can't buy everything but has to take into consideration the private budget every month. Private households have totally an average income of 1395 billion Euro per year (1998–2009) according to real data. In 1998, at that time the simulation product components: personal computer, internet and digital camera have different states of life cycle. For this reason in $t = 1998$ the stock of adopters for personal computer and internet isn't zero. Figure 3 shows the diffusion process

of product component A (PC). Product component B and C shows the same curve progression with other data.

In the starting point of simulation, there are different investigation events. Beside the reference curve (base run), there are two scenarios. The background of the first scenario is that the average willingness to pay for all intelligent applications represents about 5000 Euro (Szuppa 2007). This amount is not an ideal solution of tangible assets and services of an intelligent house for private consumption. It is called a minimum. The second scenario visualises the comfort solution with a price of 15,000 Euro (Szuppa 2007). In comparison with the changes in willingness to pay there are fluctuations in the diffusion process and the disposable income of private

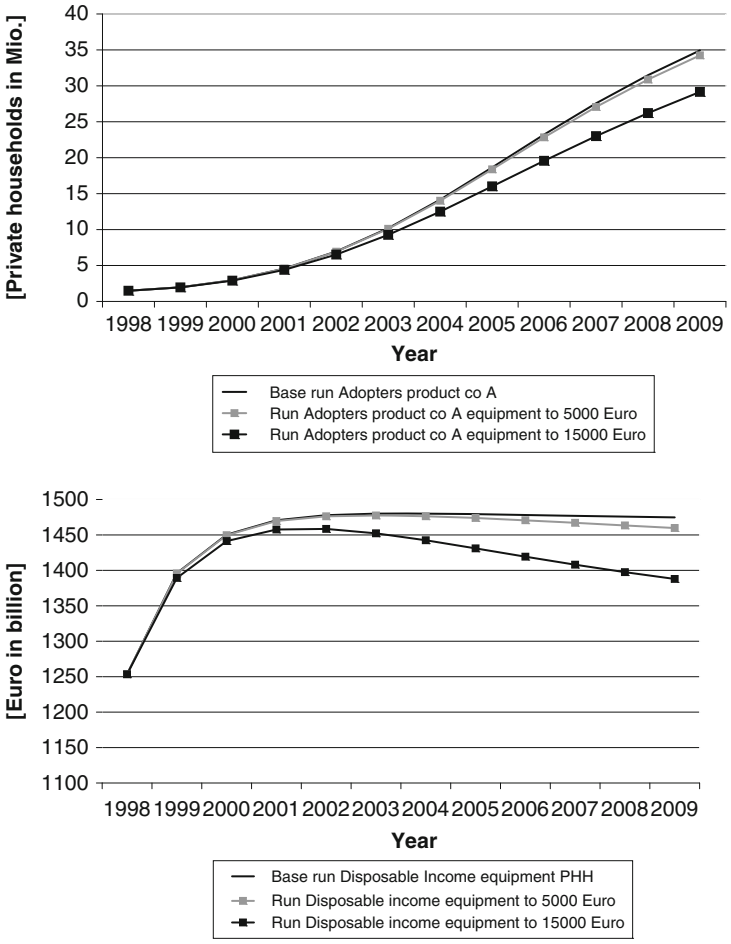


Fig. 3 Development of the diffusion process for product component A (PC) of the CoPS and income of private households - base run and two scenarios

households. How strong the changes are, depends on the definition of parameter values.

6 Summary

Complex products and systems penetrate more and more in the target group of privat consumer (B2B2C). One difficult problem for firms is, that the diffusion process slows down or delays because of high product complexity, partly non-transparent presented benefit of CoPS, stepwise purchase and high investments for consumers. By means of a detailed analysis of characteristics of CoPS as well as simulating the diffusion process, it is possible to improve the understanding about diffusion barriers by CoPS and to generate specific policies. Up to now the amount of diffusion models is very rare which analyses two interactions parallel. Furthermore the most diffusion models consider the supplier and not the consumer perspective.

References

Bass, F. M. (1969). A new product growth for model consumer durables. *Management Science, 15*(5), 215–227.

Bayus, B. L., Kim, N., & Shocker, A. D. (2000). Growth models for multiproduct interactions – Current status and new directions. In V. Mahajan, E. Muller, & Y. Wind (Eds.), *New-product diffusion models* (pp. 141–163). Boston: Kluwer Academic Publisher.

Cezanne, W. (2005). *Allgemeine Volkswirtschaftslehre*. Muenchen: R. Oldenbourg.

Chandrasekaran, D., & Tellis, G. J. (2008). A critical review of marketing research on diffusion of new products. *Marshall Research Paper Series, Working Paper MKT 01-08, University of Southern California*, 38–80.

Davies, A., & Hobday, M. (2005). *The business of projects – managing innovation in complex products and systems*. Cambridge: Cambridge University Press.

Davies, A. (1997). The life cycle of a complex product system. *International Journal of Innovation Management, 1*(3), 229–256.

Hobday, M., Rush, H., & Tidd, J. (2000). Innovation in complex products and systems. *Research Policy, 29*, 793–804.

Kroeber-Riel, W., & Weinberg, P. (2003). *Konsumentenverhalten*. Muenchen: Verlag Vahlen.

Mahajan, V., & Peterson, R. A. (1985). *Models for innovation diffusion*. California: Sage Publications.

Mahajan, V., Muller, E., & Bass, F. M. (1990). New product diffusion models in marketing: A review and directions for research. *Journal of Marketing, 54*(January), 1–26.

Milling, P. M. (1974). *Der technische Fortschritt beim Produktionsprozess*. Wiesbaden: Gabler.

Peterson, R. A., & Mahajan, V. (1978). Multi-product growth models. In J. Seth (Ed.), *Research in Marketing-Volume 1* (pp. 201–232). Greenwich: Jai Press.

Rogers, E. M. (2003). *Diffusion of innovations*. New York: The Free Press.

Sandrock, J. (2006). *System Dynamics in der strategischen Planung*. Wiesbaden: Deutscher Universitaets-Verlag.

Schmidt, S. (2009). *Die Diffusion komplexer Produkte und Systeme: Ein systemdynamischer Ansatz*. Wiesbaden: Gabler.

Singh, K. (1997). The impact of technological complexity and interfirm cooperation on business survival. *Academy of Management Journal, 40*(2), 339–367.
Szuppa, S. (2007). *Marktforschung fuer komplexe Systeme aus Sach- und Dienstleistungen im Privatkundenbereich*. Hamburg: Verlag Dr. Kovac.
Tidd, J., Bessant, J., & Pavitt, K. (2005). *Managing innovation*. Hoboken: Wiley.
Trommsdorff, V., & Steinhoff, F. (2007). *Innovationsmarketing*. Muenchen: Vahlen.

Household Possession of Consumer Durables on Background of some Poverty Lines

Józef Dziechciarz, Marta Dziechciarz, and Klaudia Przybysz

Abstract Monitoring of poverty indexes and structure of household expenditures is one of diagnostic and warning tool for social politics. Well-being or poverty of selected households' groups can be illustrated not only by mentioned categories, but also by possession of consumer durables. Thus, there is implication for poverty and possession (or lack) of some consumer durable goods. This assumption is being tested in Polish reality. In the article Polish households are grouped according to chosen classification method. Ownership of selected consumer durables was vital in obtaining clusters. Moreover, in the analysis of clusters, life quality and durable goods possession, modified poverty indexes were implemented. Obtained results were confirmed that there is the significant relationship between the social poverty and consumer durable goods.

1 Introduction

The concept of poverty as a problem of economic and social development is widely discussed in numerous studies, not only in the field of social policy (see Radziukiewicz 2006). It is also one of the main issues undertaken in the strategies of the European Union. Similarly, the issue of measuring poverty, application of appropriate measures and indicators is the subject of numerous publications, including statistical ones (see Kot 2000). The primary source of knowledge about poverty is household survey carried out regularly in most countries.

The research presented in this article was undertaken basing on the statement that the formulas of all aggregate measures of poverty are based on the choice of poverty line. Thus, the question of construction of poverty lines is a fundamental problem of identifying the poor. Furthermore we assume that the measurement of poverty should consider households' durables possession. It may have an impact on the subjective view of the financial situation of the household. It seems obvious that

K. Przybysz (✉)
University of Economics, Wrocław, Poland
e-mail: klaudia.przybysz@ue.wroc.pl

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_81,

a household which has no durable goods is in worse situation that the one which has the same level of income and possess washing machine, TV set or a car.

2 The Method

Poverty lines used in social policy are different in its nature. It is possible to apply the following classification of poverty lines (see Radziukiewicz 2006, pp. 20–21):

- Relative – defined in relation to the overall distribution of income or consumption in the country,
- Absolute – based on estimates of the cost of basic nutritional needs plus the amount necessary to meet the needs of some non-food products,
- Alternative – subjectively defined by the respondents,
- Objective – all which have not been constructed on the background of the respondents opinions.

Poverty lines and their levels in Poland in 2007 (monthly income) are as follows:

- The statutory limit of poverty – 461 PLN (less than 100 €),
- Social minimum – 820,6 PLN (about 175 €),
- Relative poverty line – 479 PLN (about 100 €),
- Subjective poverty line – ?,
- Subsistence level – 387 PLN (about 85 €).

In order to confirm the hypothesis the data on 4,941 households from the surveys conducted in 2007 was used. It was published in *Social Diagnosis 2007. Life Standard and Life Quality of Polish Society*. Initially households were classified according to income. Afterwards the classification was made according to the adopted definition of poverty lines. The test group showed no differentiation at three out of five criteria:

- The statutory limit of poverty,
- Subsistence level,
- Relative poverty line.

The next stage of the study was to determine differences or similarities in the possession of consumer durables in determined types of households using three variables (poverty measurement):

- Subjective,
- Objective,
- Composite.

Taken the results of described classifications into account *a new composite indicator* was defined. It includes a social minimum as poverty line and subjective sense of belonging to a group of poor. We can say that it is a new kind of modified poverty

line. The next stage of research was to check whether the obtained classes are differentiated according to the possession of durable goods. Due to the large variety of goods referred to as durables, it was necessary to divide the durables using some criterion. Based on the theory of multi-dimensional treatment of goods, proposed by J. Kramer (see Kramer 1993, pp. 161–164), the following classification of durable goods was used:

- Basic durable goods available to almost every household – more than 50% of households possess them,
- Standard durable goods available to 15% – 50% of households possess them,
- Luxury durable goods available to less than 15% of households.

Three groups of durable goods were established as a result of described classification (see Tables 1–3).

Figures 1–3 illustrate possession of durable goods in households divided into poor and not poor. The term *poverty* is understood as the situation of the households with monthly income lower than 820 PLN (about 175 €) and *no poverty* – with monthly income higher than this amount. The scale 0–4 or 0–5 used in following figures illustrate the number of possessed goods from groups: basic, standard or luxury accordingly. For example in the Fig. 1 the fifth column shows that over 20% of poor households possess two out of four basic durable goods (see Table 1).

As you can see, the smallest variation in discrete groups of households because of durable goods possession classification is in the field of luxury goods. In the next

Table 1 Basic durable goods in numbers and percentage

Basic durable goods	Number	Percentage
Wash machine	4 262	83,12
Car	2 883	56,30
DVD	2 557	50,18
PC	2 449	50,10

Table 2 Standard durable goods in numbers and percentage

Standard durable goods	Number	Percentage
Microwave oven	2 151	41,89
Cable TV	1 718	33,35
Satelite TV	1 375	26,65
LCD TV Set	1 016	19,77

Table 3 Luxury durable goods in numbers and percentage

Basic durable goods	Number	Percentage
Home cinema	658	12,65
Garden plot	653	12,71
Laptop, notebook	469	9,03
Dish machine	389	7,45
Summer house	205	3,80

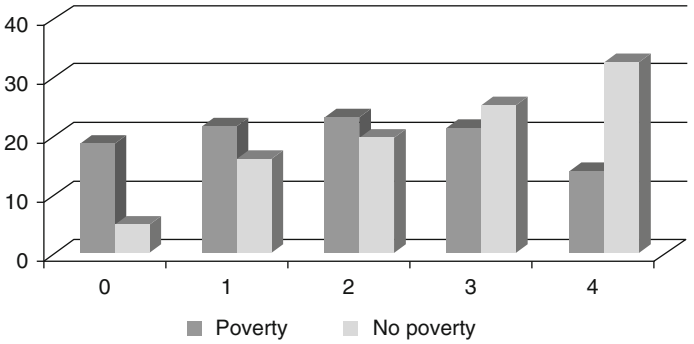


Fig. 1 Percentage of poor and not poor households possessing from 0 to 4 basic durable goods

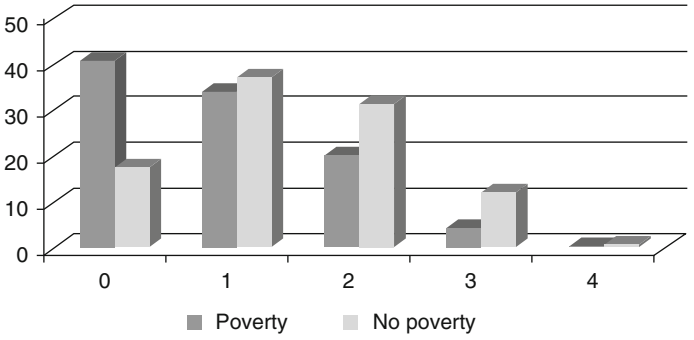


Fig. 2 Percentage of poor and not poor households possessing from 0 to 4 standard durable goods

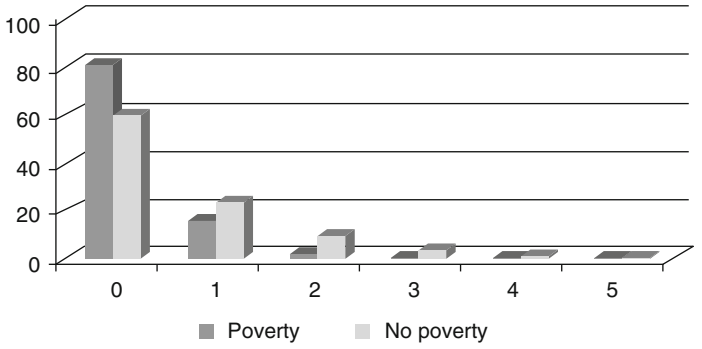


Fig. 3 Percentage of poor and not poor households possessing from 0 to 5 luxury durable goods

phase of the study households were divided according to subjective poverty line. The subjective poverty line was created on the basis of households' responses to the question whether they hardly or easily makes ends meet. As a result of these classification two groups of households were distinguished. About 77% of the respondents

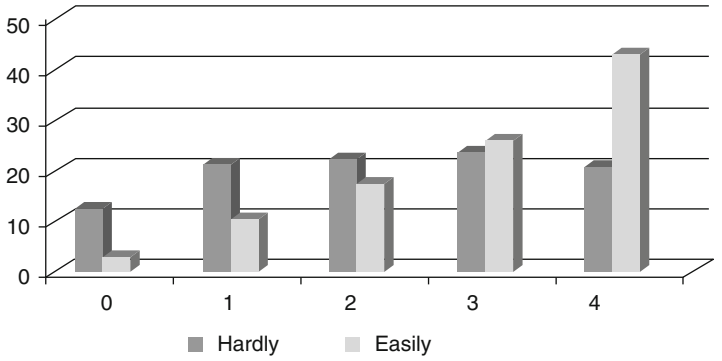


Fig. 4 Percentage of households easily and hardly making ends meet, which possess from 0 to 4 basic durable goods

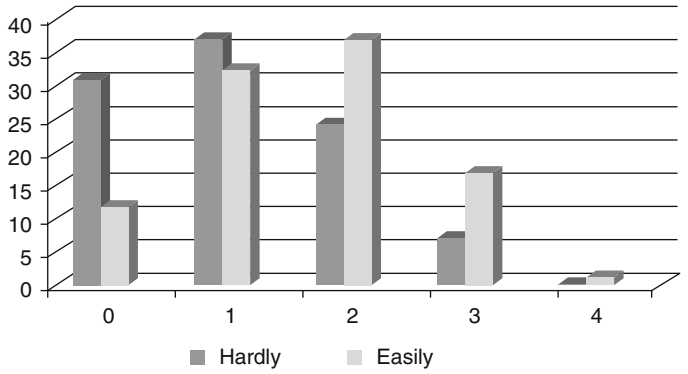


Fig. 5 Percentage of households easily and hardly making ends meet, which possess from 0 to 4 standard durable goods

with difficulty makes ends meet, the remaining 23% has no problems in this area. The following figures illustrate the households' possession of durable goods in two created classes (see Figs. 4–6).

The next stage of the study consisted in households' classification using the modified poverty line (composite indicator variable). As a result of this division three groups of households were obtained. Confirmed poverty – with income below the poverty line and the sense of poverty (27.6%), unconfirmed poverty – with income above the poverty line and the sense of poverty (50.1%), lack of poverty – with income above the poverty line and the lack of the sense of poverty (22.3%).

The distribution of households according to the new, modified poverty line, allowed to observe the possessions of households which describe its financial position as poor even though according to objective poverty line they are not in the area of poverty. Possession of three kinds of durable goods in elaborated groups of households is shown on Figs. 7–9.

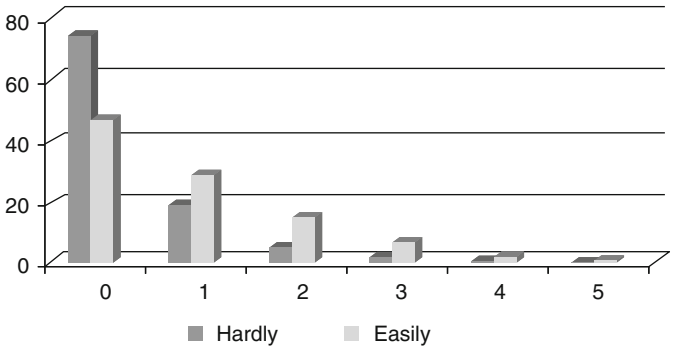


Fig. 6 Percentage of households easily and hardly making ends meet, which possess from 0 to 5 luxury durable goods

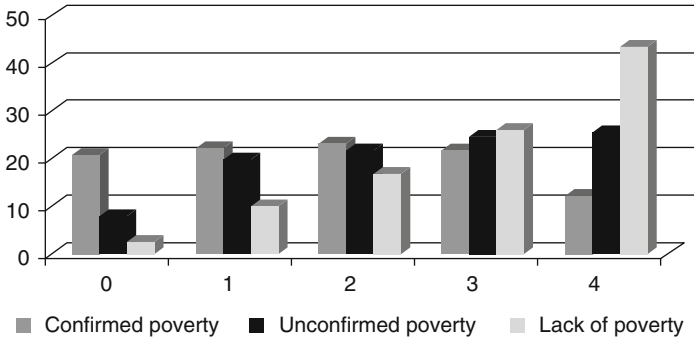


Fig. 7 Percentage of households with confirmed poverty, unconfirmed poverty and lack of poverty, which possess from 0 to 4 basic durable goods

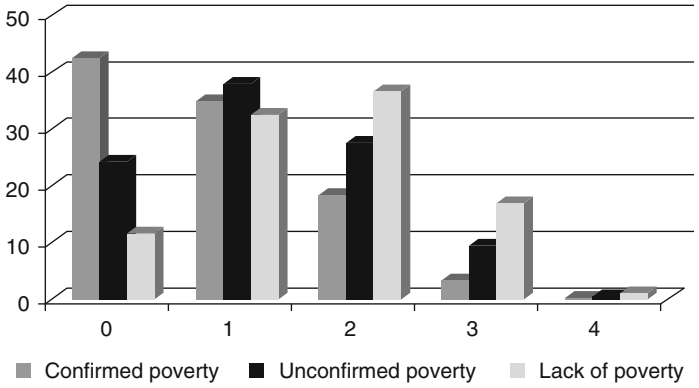


Fig. 8 Percentage of households with confirmed poverty, unconfirmed poverty and lack of poverty, which possess from 0 to 4 standard durable goods

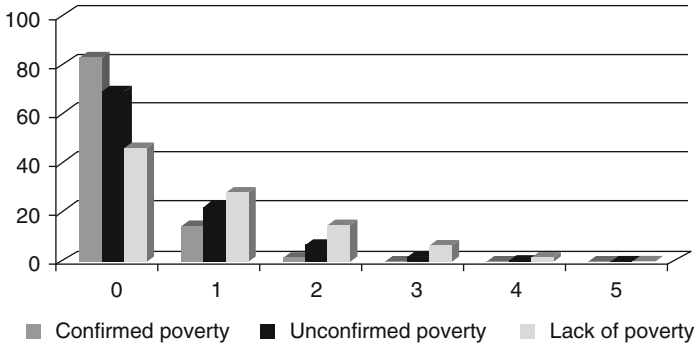


Fig. 9 Percentage of households with confirmed poverty, unconfirmed poverty and lack of poverty, which possess from 0 to 5 luxury durable goods

For each of three described divisions of households, classification trees (of discriminatory character because the dependent variable is measured on nonmetric scale (see Panek 2009) were created. A set of learner was used. In each case, the categories of variables were different: in the first case two categories – poverty; no poverty, in the second case two categories: easily; hardly, and three categories in the third case: confirmed poverty; unconfirmed poverty; lack of poverty. Other categories of variables, namely the durable goods possession, the number of persons in the household, the family head education and place of residence, remained unchanged. Analysis of the obtained results allows to conclude that respondents were well differentiated by households' possession variables. Each of the analyzed cases proved to be statistically significant. The most interesting division was brought by the application of new composite indicator. Analyzed households have been divided first in respect to the variable "basic durable goods possession", next in respect to the variable "standard durable goods possession" and at the end in respect to the variable "luxury durable goods possession". This correctness was affected only in one node of the classification tree.

3 Conclusions

The results confirm that households' durables possession have an considerable impact on the individual (subjective) assessment of the economic situation of the household. The application of the composite indicator variable allowed to extract the specific and number grup of households called *unconfirm poverty*. Further analysis of the extracted through classification trees segments may help to identify additional areas that should be taken into account in assessing the financial situation of households, in order to be able to design the most objective measures of poverty.

References

Kot, S. M. (2000). *Ekonometryczne modele dobrobytu*. Warszawa-Kraków: PWN.
Kramer, J. (1993). *Konsumpcja. Prawidłowości, struktura, przyszłość*. Warszawa: PWE.
Panek, T. (2009). *Statystyczne Metody Wielowymiarowej Analizy Porównawczej*. Warszawa: Warsaw School of Economics.
Radziukiewicz, M. (2006). *Zasięg ubóstwa w Polsce*. Warszawa: PWE.

Effect of Consumer Perceptions of Web Site Brand Personality and Web Site Brand Association on Web Site Brand Image

Sandra Loureiro and Silvina Santana

Abstract This study present a conceptual model linking web site brand personality and web site brand association to web site brand image. The model was estimated on data from consumers of online products from two countries, Spain and Scotland, using PLS technique.

1 Introduction

Brands are an important source of competitive advantage. Therefore, knowing how actual and potential clients perceive a brand is fundamental to inform its management. In brand theory, a brand is said to have attributes such as brand personality, brand association, and brand image to which brand knowledge is always linked [e.g., (Aaker 1991; Keller 1993, 1998)]. Some authors defend that the consumer-brand relationship depends largely on the successful establishment of the brand Knowledge (Keller 2003).

Brand Knowledge can be formed directly from a consumer's experience. Therefore, they might be crucial mediators between brand experience and consumer-brand relationship. If such a relation proves, understand the way these concepts interrelate with each other might be valuable to inform marketing strategy formulation, namely, in what concerns brand management.

The main goal of this paper is to test a model relating web site brand personality and web site brand association with the formation of web site brand image from an experiential view. The model was estimated on data from 195 consumers of online products from two countries, Scotland and Spain, using PLS technique.

To our best knowledge, this is the first time web site brand Knowledge (association, personality, and image) are addressed in such a way and the study differs from previous work which have related brand Knowledge of goods and services (Bart

S. Loureiro (✉)
University of Aveiro, Campus de Santiago, 3810-193 Aveiro, Portugal
e-mail: sandra.loureiro@ua.pt

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_82,

et al. 2005; Chang and Chieng 2006), sold through virtual stores (web site) or physical stores. Second, this study focuses on consumers experiences in two European countries with very different levels of Internet use for shopping. Given the paucity of cross-country studies in this area, using PLS might prove to be valuable to considerably advance existing knowledge and enhance current practices of web use for retailing.

2 Theoretical Background and Hypotheses

In this study we state that web site brand personality, web site brand association and web site brand image all hold different information that link to a web site brand (Aaker 1991). Furthermore, we assume that web site brand personality and web site brand association are important determinants of website image.

Brand image is defined here as perceptions about a brand as reflected by the brand associations held in consumer memory (Keller 1993). Brand personality is defined as the set of human characteristics associated with a brand (Aaker 1997). It is a comprehensive concept, which includes all the tangible and intangible traits of a brand, such as beliefs, values, prejudices, features, interests, and heritage. A brand personality makes it unique. Brand personality is seen as a valuable factor in increasing brand engagement and brand attachment, in much the same way as people relate and bind to other people. Researchers have proposed that brand personality is an aspect of brand image (Keller 1993, 1998; Plummer 2000) and results from empirical studies indicate that brand personality have a statistically significant positive influence on brand image (O'Cass and Lim 2001).

According to previous studies (Chang and Chieng 2006; Keller 1998), brand association is defined as the information linked to the node in memory. This information reflects an association between a range of aspects and the brand in the mind of the consumer. Brand associations have been presented as critical components in developing a brand image (Keller 1993) and empirical studies have shown that brands associations lead to the formation of a distinct brand image in the minds of consumers (Hsieh 2002). In this work, the above three concepts are transposed to the context of web site brand and we hypothesize that:

H1: Web site brand personality significantly and positively influences web site brand image

H2: Web site brand association significantly and positively influences web site brand image

3 Methods

The surveys were conducted in June 2008 through face-to-face interviews in universities of Spain and Scotland. We collected 95 completely filled questionnaires from students of Spain and 100 from students of Scotland. Each sub-sample has the

same average age of 24 years. The respondents were split almost equally in terms of gender for both countries.

Brand association was measured using two dimensions (product and organization) (Barclay et al. 1995; Carmines and Zeller 1979). Brand personality was operationalized using 5 dimensions (sincerity, excitement, competence, sophistication, and ruggedness) (Aaker 1997) and brand image with 3 dimensions (function, experience, and symbolic) (Chang and Chieng 2006; Keller 1993). Each statement of the questionnaire was recorded on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree).The questionnaire was built in English, translated to Spanish and than translated back to English in order to guarantee the content validity.

A structural equation model approach using Partial Least Squares (PLS) (Ringle et al. 2005) was employed to test the hypotheses of this study. PLS (Partial Least Squares) is based on an iterative combination of principal components analysis and regression, and it aims at explaining the variance of the constructs in the model. In terms of advantages, PLS simultaneously estimates all path coefficients and individual item loadings in the context of a specified model, and as a result, it enables researchers to avoid biased and inconsistent parameter estimates. Based on recent developments (Chin et al. 2003), PLS has been found to be an effective analytical tool to test interactions by reducing type II error. By creating a latent construct which represents the interaction term, a PLS approach significantly reduces this problem by accounting for the error related to the measures. In fact, PLS models are based on prediction-oriented measures, not covariance fit like covariance structure models developed by Karl Jöreskog (or LISREL program developed by Jöreskog and Sörborn).

LISREL estimates causal model parameters aiming at minimizing the discrepancies between the initial empirical covariance data matrix and the covariance matrix deduced from the model structure and the parameter estimates (Barclay et al. 1995). PLS seeks to maximize variance explained in constructs and/or variables, depending on model specification. In addition, LISREL offers a number of measures of overall model "fit" such as the X^2 goodness-of-fit (which are related to the ability of the model to account for the sample covariances). PLS does not possess these kind of overall fit measures, relying instead on variance explained (i.e. R^2) as an indicator of how well the technique has met its objective (Barclay et al. 1995). In spite of that, there are several fit indices available on PLS software (Ringle et al. 2005) such as communality and redundancy measures and Stone-Geisser's Q^2 measure, which can be used to evaluate the predictive power of the model.

As a substitute to parametric global goodness of fit measures that are used in LISREL technique (Tenenhaus et al. 2005) proposes the geometric mean of the average communality (outer model) and the average R^2 (inner model) (going from 0 to 1) as overall goodness of fit (GoF) measures for PLS (Cross validated PLS GoF):

$$GoF = \sqrt{\overline{communality} \cdot \overline{R^2}} \tag{1}$$

Table 1 Measurements Results

Manifest Construct	LV Index Values	Item Loading	Composite reliability	AVE*
		Spain		
Brand association	3.5		0.8441	0.7309
AS1:Product		0.9000		
AS2:Organization		0.8070		
Brand personality	3.2		0.8769	0.7039
PS1:Excitement		0.8490		
PS2:Sophistication		0.8660		
PS3:Ruggedness		0.8000		
Brand Image	3.3		0.8702	0.6915
IS1:Functional		0.7650		
IS2:Experience		0.8820		
IS3:Symbolic		0.8440		
		Scotland		
Brand association	3.6		0.9542	0.9125
ASc1:Product		0.9550		
Asc2:Organization		0.9550		
Brand personality	3.2		0.8780	0.7063
PSc1:Excitement		0.8850		
PSc2:Sophistication		0.7870		
PSc3:Sincerity		0.8460		
Brand Image	3.2		0.8834	0.7164
ISc1:Functional		0.8290		
ISc2:Experience		0.8400		
ISc3:Symbolic		0.8700		

*AVE Average Variance Extracted.

4 Results

A PLS model should be analyzed and interpreted in two stages. First, the adequacy of the measures (see Tables 1 and 2) is assessed by evaluating the reliability of the individual measures and the discriminant validity of the constructs (Hulland 1999). Then, the structural model is evaluated.

All the loadings (Table 1) of reflective constructs approach or exceed 0.707, which indicates that more than 50% of the variance in the manifest variable is explained by the construct (Carmines and Zeller 1979), excepting for the construct brand personality. Sincerity and competence were eliminated from the Spanish sample. Competence and ruggedness were eliminated from the Scottish sample. Composite reliability was used to analyze the reliability of the constructs since this has been considered a more exacting measurement than the Cronbach's alpha (Fornell and Larcker 1981).

Table 1 indicates that all constructs are reliable since the composite reliability values exceed the threshold of 0.7 and even the strictest one of 0.8 (Nunnally 1978). The measures also demonstrates convergent validity as the average variance of manifest variables extracted by constructs (AVE) is at least 0.5, indicating that more variance is explained than unexplained in the variables associated to a given construct.

To assess discriminant validity the square root of AVE should be greater than the correlation between the construct and other constructs in the model (Fornell and Larcker 1981). Table 2 shows that this criterion has been met.

The structural results for Spain are presented in Fig. 1.

All the path coefficients are found to be significant at the 0.001 level and all the coefficients signs are in the expected direction. In this study a nonparametric approach, named Bootstrap, was used for estimating the precision of the PLS estimates and support the hypotheses. Accordingly, 500 samples sets were created in order to obtain 500 estimates for each parameter in the PLS model. Each sample was obtained by sampling with replacement to the original data set (Chin 1998; Fornell and Larcker 1981). The results of this procedure support all the hypothesized relations. The model also demonstrates a high level of predictive power (R^2) as the modeled constructs explains 66.6% of the variance in the brand image. The overall goodness of fit (Tenenhaus et al. 2005) reveals a good fit.

Table 2 Discriminant Validity: square root of AVE and correlations of constructs

Construct	Correlations of constructs		
	Brand association	Brand Image	Brand personality
	Spain		
$AVE^{1/2}$	0.8549	0.8316	0.8390
Brand association	1.0000	0.7663	0.4448
Brand Image	0.7663	1.0000	0.5926
Brand personality	0.4448	0.5926	1.0000
	Scotland		
$AVE^{1/2}$	0.9553	0.8464	0.8404
Brand association	1.0000	0.7324	0.6732
Brand personality	0.7324	1.0000	0.6845
Brand Image	0.6732	0.6845	1.0000

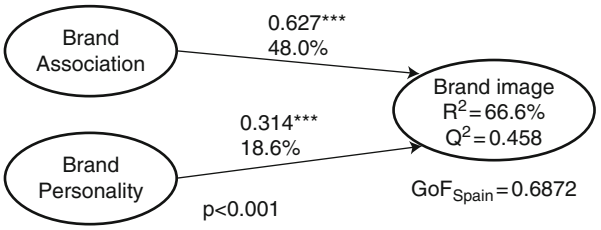


Fig. 1 Structural results (Spain)

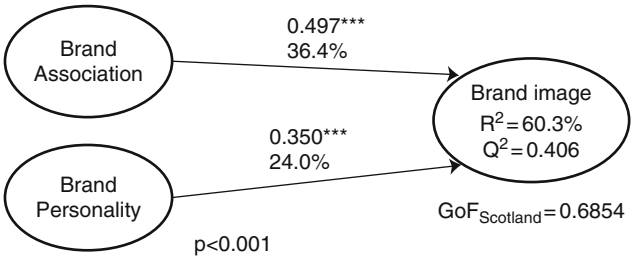


Fig. 2 Structural results (Scotland)

The multiplication of the Pearson correlation value for the path coefficient value of each two constructs reveals that 48.0% of the brand image variability is explained by the brand association.

The structural results for Scotland are presented in Fig. 2.

All the path coefficients are significant at the 0.001 level and all the coefficients signs are also in the expected direction. Like in the Spanish sample, the Bootstrap approach with n = 500 was used and all the hypothesized relations were supported. The value of Q^2 is positive attesting that the relations in the model have predictive relevance. The model also demonstrates a high level of predictive power, as the modeled constructs explains 60.3% of the variance in the brand image and the overall goodness of fit proposed also reveals a good fit. The multiplication of the Pearson correlation value for the path coefficient value of each two constructs reveals that 39.4% of the brand image variability is explained by the brand association and that 24.0% is explained by the brand personality.

Finally, the differences between the Scottish and the Spanish samples are compared using a t-test of m + n + 2 degrees of freedom (m = Spain sample size and n = Scotland sample size). This test uses the path coefficients and the standard errors of the two structural paths calculated by PLS with the samples of both countries, using the following expression:

$$t = \frac{(\beta_{\text{Spain}} - \beta_{\text{Scotland}})}{\text{S}_\text{p} \times \sqrt{\left(\frac{1}{\text{m}} + \frac{1}{\text{n}}\right)}} \tag{2}$$

$$S_p = \sqrt{\left[\frac{(\text{m}-1)^2}{(\text{m}+\text{n}-2)} \times SE^2_{Spain} + \frac{(\text{n}-1)^2}{(\text{m}+\text{n}-2)} \times SE^2_{Scotland}\right]}$$

The t-test results (Table 3) show that there are not statistically significant differences between the two countries in any of the two structural paths (at critical $t - value = |1.960|$).

Table 3 Multi-group Analysis Results

Structural paths	Standard error Spain	Standard error Scotland	Sp[1]	$\beta_{spain} - \beta_{scotland}$	t-test
Brand association → Brand image	0.0974	0.0922	0.9306	0.1299	0.0014
Brand personality → Brand image	0.0898	0.0983	0.9273	−0.0362	−0.0004

[1]Unbiased estimator of average error standard variance

5 Discussion

This study is the first attempt to considerer the web site brand in a structural model using PLS approach, which analyze simultaneously the causal order between web site brand association and web site brand image and between web site brand personality and web site brand image.

The results show that web site brand association and web site brand personality are good predictors of web site brand image and that the two hypotheses are confirmed for the Scottish and the Spanish samples. However, competence and ruggedness were eliminated from the Scottish sample and sincerity and competence were eliminated from the Spanish sample. The differences between the two countries are not statistically significant but the way students see web site brand personality might depend on the country culture. This area should be the object of further investigation.

Traditionally, brand image and brand personality are different constructs. However, PLS technique seems to evidence some correlation between the competence (eliminated in this analyze) of brand personality and the symbolic part of brand image.

Further directions for future work have been pointed out by this first study of web site brand knowledge. The model is being redesigned to include other constructs and we are planning to extend our research to other countries, such as Brazil, USA, Germany, Portugal and Poland. With a cross-country approach we will be able to analyze the impact of culture on consumers' perception and test the effect of globalization, advancing existing knowledge and generating valuable information to decision makers, marketers and web designers.

References

Aaker, D. (1991). *Managing brand equity: Capitalizing on the value of a brand name.* New York: The Free Press.

Aaker, J. (1997). Dimensions of brand personality. *Journal of Marketing Research, 34*, 347–356.

Barclay, D., Higgins, C., & Thompson, R. (1995). The partial least squares (PLS) approach to causal modeling, personal computer adoption and use as an illustration. *Technology Studies, 2*, 285–309.

Bart, Y., Shankar, V., Sultan, F., & Urban, G. (2005). Are the drivers and the role of online trust the same for all web sites and consumers? A large-scale exploratory empirical study. *Journal of Marketing, 69*, 133–152.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment.* London: Ed. Sage Publications, Inc.

Chang, P-L., & Chieng, M-H. (2006). Building consumer-brand relationship: A cross-cultural experiential view. *Psychology and Marketing, 23*(11), 927–959.

Chin, W. (1998). The partial least squares approach to structural equation modeling. In G. A. Marcoulides (Ed.), *Modern methods for business research* (pp. 295–336). Mahwah, NJ: Lawrence Erlbaum Associates Publisher.

Chin, W., Marcolin, B. L., & Newsted, B. L. (2003). A partial least squares latent variable modelling approach for measuring interaction effects: Results from a Monte Carlo simulation study and an electronic mail emotion/adoption study. *Information Systems Research, 14*(2), 189–217.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural models with unobservables variables and measurement error. *Journal of Marketing Research, 28*, 39–50.

Hulland, J. (1999). Use of partial least squares (PLS) in strategic management research: A review of four recent studies. *Strategic Management Journal, 20*(2), 195–204.

Hsieh, M. (2002). Identifying brand image dimensionality and measuring the degrees of brand globalization: A cross-nation study. *Journal of International Marketing, 10*, 46–67.

Keller, K. (1993). Conceptualizing, measuring and managing customer-based brand equity. *Journal of Marketing, 57*, 1–22.

Keller, K. (1998). Strategic brand management: Building, measuring and managing brand equity. New York: Prentice Hall.

Keller, K. (2003). Brand synthesis: The multidimensionality of brand knowledge. *Journal of consumer research (March), 29*, 595–600.

Nunnally, J. (1978). *Psychometric theory* (2nd ed). New York: McGraw-Hill.

O'Cass, A., & Lim, K. (2001). The influence of brand association on brand preference and purchase intention: As Asia perspective on brand association. *Journal of International Consumer Marketing, 14*, 41–71.

Plummer, J. (2000). How personality makes a difference. *Journal of Advertising Research, 40*(6), 79–83.

Ringle, C. M., Wende, S., & Will, A. (2005). SmartPLS 2.0 (beta), www.smartpls.de, Hamburg.

Tenenhaus, M., Vinzi, V. E., Chatelin, Y.-M. & Lauro, C. (2005). PLS path modeling. *Computational Statistics & Data Analysis, 48*, 159–205.

Perceptually Based Phoneme Recognition in Popular Music

Gero Szepannek, Matthias Gruhne, Bernd Bischl, Sebastian Krey, Tamas Harczos, Frank Klefenz, Christian Dittmar, and Claus Weihs

Abstract Solving the task of phoneme recognition in music sound files may help for several practical applications: it enables lyrics transcription and as a consequence could provide further relevant information for the task of an automatic song classification. Beyond it can be used for lyrics alignment e.g. in karaoke applications.

The effect of both different feature signal representations as well as the choice of the appropriate classifier are investigated. Besides, a unified `R` framework for classifier optimization is be presented.

1 Introduction

This work is an extension of previous studies at the Fraunhofer IDMT on automatic phoneme classification in polyphonic music (Gruhne et al. 2007). An accurate phoneme recognition in music may yield a basis for several applications like automatic lyrics extraction (and further automatic classification of songs) as well as for the automatic alignment of previously known lyrics to music for karaoke applications. The specific goal of this work consists of the examination of different feature sets extracted from audio data combined with an appropriate choice and parameter tuning of classifiers. Concerning the feature sets, perceptive phenomena have been more and more introduced in audio processing during the last years. It is of interest up to what extent it is beneficial to model human sound processing. A detailed neurophysiological simulation model of the human auditory periphery (serving as basis for feature extraction) is compared to a simpler and computationally less expensive phenomenological one. Auditory model-based features are opposed to well-known standard acoustical feature vector representations. A unified framework for the statistical programming language `R` is presented that easily allows to tune, optimize and compare the influence of different classifiers for specific data situations and the

G. Szepannek (✉)
Faculty of Statistics, Dortmund University of Technology, D-44221 Dortmund, Germany
e-mail: szepannek@statistik.tu-dortmund.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_83,

given task. A description of the task in Sect. 2 is followed by a brief introduction to auditory modelling in Sect. 3. Feature extraction from audio data (based on the original waveform as well as on the auditory simulation model output) is described in Sect. 4. The framework for classifier optimization is presented in Sect. 5. Finally, the results of the study, a discussion and a summary are given in Sects. 6 and 7.

2 Description of the Task

The data under investigation consists of 45 files of popular music (30 male and 15 female singers, 44.1 kHz, e.g. *I'm a believer* (Monkees), *Sweet dreams* (Eurithmics), *Zepher* (Red Hot Chili Peppers), *Billie Jean* (Michael Jackson), *Killing me softly* (Roberta Flack), *Song* #2 (Blur), ...). The music files are split into training and test files (2/3 : 1/3). All songs are phonetically manually labelled at the Fraunhofer IDMT according to the TIMIT phonetic transcription. Only one single feature vector (from a 1,024 samples window, i.e. 23.6 ms) is computed per phoneme to avoid highly correlated observations. In automatic speech recognition monophones are typically modelled as three state hidden Markov models (HMMs) where the second state corresponds to its stationary part (see e.g. Gold and Morgan 2000, p. 365). We assumed to maximise the chance to hit this "inner" steady state of the phonemes when considering the window at half of the phonemes total duration. Fifteen different vowels as well as consonants were taken into the analysis as far as there were at least 50 observations of each phoneme in total. The resulting classes are: */a/, /ae/, /e/, /ee/, /i/, /j/, /l/, /m/, /n/, /o/, /oa/, /oe/, /ou/, /r/ and /w/*. Finally, there were 1,549 phonemes in the training and 672 phonemes in the test set. For the detailed auditory model the `*.wav` files have to be amplitude normalized before processing to be able to set the absolute sound pressure level (SPL) (see also Sect. 3).

We applied sinusoidal preprocessing as it turned out to be beneficial (Gruhne et al. 2007). Basically, the audio signal is considered to be a sum of voice and background. The voiced part is further modelled as a sum of sinusoids of the estimated fundamental and its harmonics. The amplitudes of all other fourier frequencies are set to 0 in the spectral domain and the result is back-transformed to the time domain (for further details, see Gruhne et al. 2007).

3 Auditory Modelling

Several well-known psycho acoustical phenomena can be traced back to sound processing in the auditory system, e.g. nonlinear frequency resolution and amplitude saturation or masking effects. Basically, the sound wave is nonlinearly bandpass-filtered at the inner ear along the *basilar membrane* (BM) and transduced into electric impulses (action potentials, APs) at the *auditory nerve fibres* (ANFs) of different *center frequency* (CF) by *inner hair cells*. A simple computational auditory

model (referred to as "Seneff-model", Slaney 1998) phenomenologically imitates human auditory sound processing within a chain of five successive steps, consisting of: critical band filtering *(BM excitation)*, halfway rectification saturating non linearity *(inner hair cell current)*, short term adaption circuit *(synaptic neurotransmitter release)*, low pass filtering *(nerve fibre: synchrony reduction)* and rapid automatic gain control *(nerve fibre: refractory effect)*. The output of the model can be interpreted as *time varying neural firing rates* at 40 different ANFs (of 0.5 bark CF difference).

Besides this, also a very detailed and computationally more intensive model of the human auditory periphery is implemented where all steps reproduce neurophysiological measurements. It simulates exact firing times of 251 different ANFs with CF differences of 0.1 bark (Szepannek et al. 2005).

Figure 1 (left) shows the average auditory nerve firing activity of the detailed model during 200 ms along the BM (abscissa) for a 1 kHz sine of different sound pressure level (SPL, ordinate). A level of 0 dB SPL denotes the threshold of hearing (Gold and Morgan 2000). Figure 1 (right, bottom) shows the response of the auditory simulation model to some vowel /a/. The ordinate represents the unrolled inner ear (BM) while the abscissa denotes the time. The output of the simulation model is binary and of the form

$$X_i(t) = \begin{cases} 1, \text{ AP of ANF } i \text{ at time } t, \\ 0, \text{ else.} \end{cases}$$

It can be seen that different positions along the BM are differentially excited (according to the signal frequencies). The ANFs further respond periodically with the signal period. This phenomenon is commonly referred to as *phase locking* (Szepannek et al. 2009). For the studies in this paper, 50 repetitive simulations of the ANFs of different type are pursued. The signals were presented to the auditory model at a (typical) level of 62.5 dB SPL (Szepannek et al. 2009).

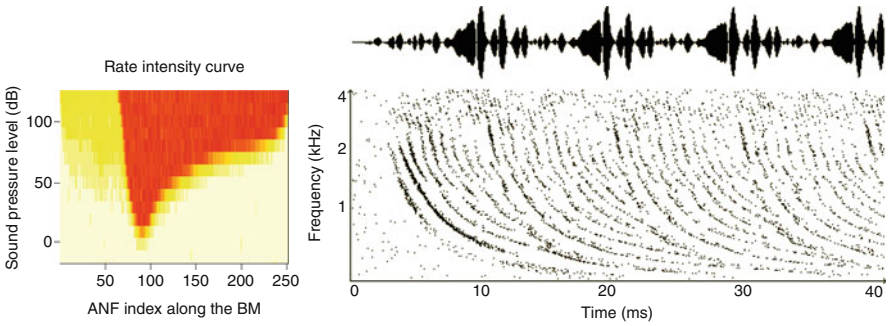


Fig. 1 Tuning curve for a 1,000 Hz sine sound of different sound pressure level (*left*) and output of the auditory model for a vowel /a/ (*right*)

4 Feature Extraction

A key idea of timbral feature extraction is the *source-filter model* (of speech production). Speech signal waves are excited at the glottis (either noisy or periodic) and get their characteristic timbre being filtered by the specific shape of the vowel tract. Thus, the filter coefficients of (fixed) order p meaningfully represent the sound characteristics. These *linear predictive (filter) coefficients (LPCs)* are derived by Levinson-Durbin recursion to minimize the predictive error (see e.g. Gold and Morgan 2000). According to former studies (Szepannek et al. 2008) a choice of $p = 16$ is used here.

Based on the principles of neural information coding mentioned above two different non-standard feature sets are extracted from the simulated auditory neural response (see Sect. 3). *Place / mean rate features* (MR) count the neural activity at different ANFs independently of its temporal fine structure, i.e.

$$X_i^{MR} = \sum_{t \in \text{window}} X_i(t)/\text{window size}.$$

According to Allen (1994) groups of eight neighbouring ANFs of the detailed auditory model in the CF range of [200, 6400] Hz are averaged to build a 24 dimensional feature vector.

On the other hand, *average localized synchrony detection* features (ALSD, Ali et al. 2002) temporally encode neural auditory information:

$$X_k^{ALSD} = \frac{1}{3} \sum_{l=k-1}^{k+1} A_s \tan^{-1}\left[\frac{1}{A_s}\left(\frac{\langle|X_l^{PSTH}(t) + X_l^{PSTH}(t-n_k)|\rangle - \delta}{\langle|X_l^{PSTH}(t) - \beta^{n_k} X_l^{PSTH}(t-n_k)|\rangle}\right)\right] \tag{1}$$

with $X_l^{PSTH}(t)$ being the time-varying firing rate of ANF l (estimated by the post stimulus time histogram of the neural activity in time bins of $\frac{1}{14700}$ s averaged over all simulations and eight neighbour ANFs as for X^{MR} for the detailed model). The $\langle . \rangle$ operator denotes temporal averaging, n_i is the period (in time bins) of the CF of ANF i. Basically, the denominator checks, whether on average the neural activity is the same as it has been one (CF-)period before. The constant $\beta = 0.99$ avoids obtaining zeros in the denominator. $\delta = 60\,\text{spikes}\,dt\,s^{-1}$ corrects for spontaneous neural activity and $A_s = 4$ is a scaling constant. According to (1) the X^{ALSD} representation consists of a 22 dimensional feature vector.

Mel frequency cepstral coefficients (MFCCs) (see e.g. Gold and Morgan 2000, pp. 280–288) have recently become popular for speech and music analysis. They also rely on the source-filter model of speech production: in the spectral domain the signal is the product of the excitation and the filter amplitudes (of the vowel tract). Building the logarithm changes this into a sum. A subsequent inverse discrete fourier transform can be interpreted as a "spectral analysis of the log-spectrum": strong periods in the spectrogram represent the fundamental and its harmonics and are captured in the higher coefficients (*quefrencies*) as well as noise is. The

characteristic shape of the log-spectrum is represented in the lower coefficients. Thus, only the lowest q coefficients are used for further timbre analysis. In this application, a typical value of $q = 13$ is chosen. To imitate human perception frequency grouping according to the mel scale is performed. The log transform can be further compared to human auditory nonlinear amplitude saturation (Szepannek et al. 2009).

Perceptual linear prediction coefficients (PLPs) also take into account human auditory sound processing (see Gold and Morgan 2000, p. 299). Before computing LPCs (see above) the sound signal is transformed into the frequency domain where amplitudes are compressed (typically by building cubic roots) and frequencies are grouped according to the perceptive mel scale. After some inverse back-transform into the time domain, LPCs are calculated. Also, an order of $p = 16$ is chosen for this work (Szepannek et al. 2008). For standard features like MFCCs, LPCs and PLPs an `R` implementation of the Matlab `rastamat` toolbox (Ellis 2005) is used.

5 Classifier Tuning

The aim of this work is to investigate the combination of both feature extraction and the choice of an adequate classifier. There exist numerous different classification algorithms (for an overview see e.g. Hastie 2001), many requiring the choice of additional free parameters. The list below shows the classifiers that were implemented for this study (in brackets the parameters that were varied): *SVMs with polynomial kernels* ($K(x, y) = (1 + \langle x, y\rangle)^d$, **PSVM**, $d \in \{1, 2, 3, 4\}$, cost of constraints violation $c \in 2^{\{-4,-3,\ldots,3,4\}}$), *SVMs with RBF kernels* ($K(x, y) = e^{-||x-y||^2/\gamma}$, **RSVM**, $\gamma \in 2^{\{-4,-3,\ldots,3,4\}}$), cost $c \in 2^{\{-4,-3,\ldots,3,4\}}$), *linear discriminant analysis* (**LDA**, -), *quadratic discriminant analysis* (**QDA**, -), *mixture discriminant analysis* (**MDA**, equal number of subclasses $\in \{2, \ldots, 5\}$), *naive Bayes* (**NB**, -), *classification trees* (**RPART**, factor of required improvement for a split to be kept in the tree model $\in \{0.005, 0.01, 0.03\}$), *random forests* (**RF**, -) and *k nearest neighbours* (*k***NN**, $k \in \{1, 2, 3, 4, 6, 8, 10\}$). All classifiers are evaluated in `R` using the packages `kernlab`, `MASS`, `e1071`, `mda`, `rpart`, `randomForest` and `kknn`. The free parameters are optimized on grids using an internal fivefold cross validation (cv) on the training data. A typical problem using the programming language `R` for classification purposes is the heterogenity of the different implemented algorithms. A framework has been developed in order to easily enable optimizing and benchmarking different classifiers using the `R` package `{mlr}` (Bischl 2009). Its features are: an object oriented `S4` interface to `R` classification methods, easy extension to new methods, it provides a unified call of different methods, bootstrapping, cross-validation, train/test splits, parameter tuning and benchmarking of different classification algorithms are possible (e.g. by 'double cv' with tuning on an inner cv).

6 Results and Discussion

At first glance, the choice of the feature set appears to play the dominant role on the accuracies. Figure 2 (left) shows the performances as a function of the choice of the feature set. As simple LPCs show the worst results on average it turns out to be worth including perceptive phenomena into feature extraction design. No strong advantages are observed using the detailed neurophysiologically parameterized auditory model instead of the simpler phenomenological one. Concerning auditory model based feature extraction ALSD outperforms MR feature extraction. Nevertheless standard features like MFCCs and PLPs show the best results. Even their averaged accuracies over all classifiers are better than the best results for optimized classifiers for any of the auditory model based feature sets. It should be noted that these feature sets both include simplified perceptual models as well as a speech production based motivation. The auditory model based features on the other hand are restricted to modelling perception. An additional cepstral transformation of the auditory model based features did not improve the results. There is no equivalent to masking effects included in MFCC or PLP feature extraction. Apparently, this effect is not of relevance for this application. Figure 3 shows the performance of the classification algorithms compared to the average accuracy on each data set separately. The best classifier is not the same for all feature sets emphasizing importance of a problem-specific classifier tuning. Some methods (NB, RPART and kNN) are a bad choice for any of the investigated feature sets. Some other classifiers (SVMs and LDA) are often among the best methods. MDA shows good results for the non-auditorily extracted features whereas random forests only perform well for the auditory model based features.

In order to obtain general hints on the choice of the classifier, a consensus ranking is derived (see e.g. Hornik and Meyer 2007). Significant differences in test accuracy of any two classifiers for each data set are investigated using McNemar's test (Dietterich 1998). The results are summarized using a Bradley-Terry model for paired comparisons. Table 1 shows the resulting π_i that can be interpreted as "prior

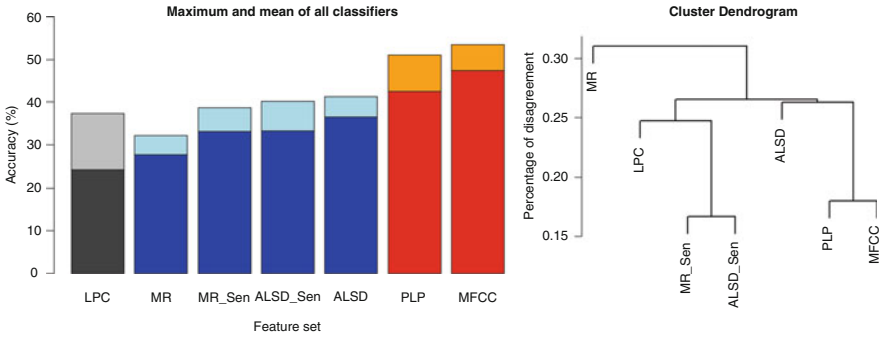


Fig. 2 Maximal/average (*light/dark grey*) accuracy over different classifiers (*left*). Feature sets clustered according to distances given by the fraction of differently predicted (test-)objects (*right*)

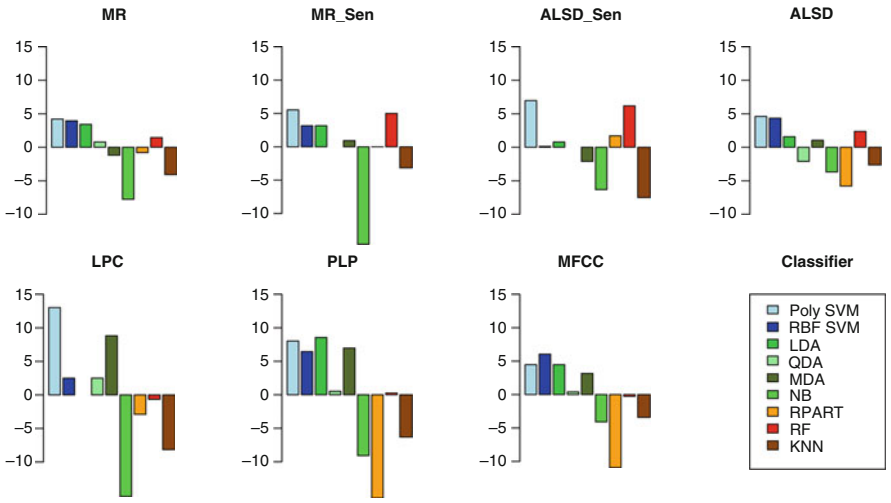


Fig. 3 Classifier performance compared to average accuracy per data feature set

Table 1 Consensus ranking of the classifiers over all data sets

	PSVM	RSVM	RF	MDA	LDA	QDA	RPART	kNN	NB
π_i	**0.424**	0.155	0.141	0.108	0.099	0.029	0.020	0.015	0.009

probabilities" for each specific classifier to be the significantly best choice. The consensus ranking of the classifiers strongly suggests the use of optimized SVMs with polynomial kernel (cf also Fig. 3). Nevertheless, the best overall results (53.42%) are obtained using RBF-kernel SVMs and MFCC features once again emphasizing the importance of problem specific classifier choice. Random forests as well as LDA and MDA also appear to be a good choice in general. Nevertheless, generalization of these results should be handled with care. The tuned parameters of the optimal model are $\sigma = 0.0625$ of the RBF kernels and complexity parameter $c = 2$. But the results strongly depend on the parameters, within the investigated parameter grid also accuracies below 20% are observed. Finally, Fig. 2 (right) identifies MFCCs and PLPs as well as the simpler auditory (Senneff) features to be more similar to each other than the other features by average linkage clustering. It should be further noted that – in contrast to modelling continuous speech the task of classifying single frames becomes more complicated, especially due to the heterogeneous (nonstationary) polyphonic background noise. The use of HMMs smoothes over successive frames for continuous modelling. The incorporation of posterior probability estimates of optimized classifiers could improve the use of standard Gaussian (MDA like) mixtures for continuous modelling (see e.g. Krueger et al. 2005). Further attention could be laid on feature combination as in Szepannek et al. (2009).

7 Summary

A task with many practical applications has been investigated: automatic recognition of phonemes in popular music. Specific interest of the study was the investigation of the influence of different feature representations in combination with the choice and tuning of the appropriate classification method. A new `R` package framework has been presented to solve the latter task. In conclusion both is beneficial: taking into account speech production as well as perception. No improvements have been observed for high degrees of precision in auditory modelling. Nonetheless, the appropriate choice and tuning of the classifier is of importance. The work could be further extended towards feature combination and modelling continuous singing.

References

Ali, A., van der Spiegel, J., & Mueller, P. (2002). Robust auditory-based speech recognition using average localized synchrony detection. *IEEE Transactions on Speech and Audio Processing, 10*(5), 279–292.

Allen, J. (1994). How do humans process and recognize speech? *IEEE Transactions on Speech and Signal Processing, 2*(4), 567–577.

Bischl, B. (2010). The `mlr` package: machine learning in `R`. http://algorithm-forge.com/bischl/mlr/

Dietterich, T. (1998). Approximate statistical tests for comparing supervised classification algorithms. *Neural Computation, 10*, 1895–1923.

Ellis, D. (2005). *PLP and RASTA (and MFCC, and inversion) in Matlab.* from http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/.

Gold, B., & Morgan, N. (2000). *Speech and audio signal processing*. NY: Wiley.

Gruhne, M., Schmidt, K., & Dittmar, C. (2007). Detecting phonemes within the singing of polyphonic music. In *Proceedings of the International Conference on Music Communication Science (ICOMCS)*, Dec. 5–7, Sydney, Australia.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New Jersey: Springer.

Hornik, K., & Meyer, D. (2007). Consensus rankings from benchmarking experiments. In R. Decker, H. Lenz and W. Gaul (Eds.), *Advances in data analyses* (pp. 163–170). Heidelberg: Springer.

Krueger, S., Schafföner, M., Katz, M., Andelic, E., & Wendemuth, A. (2005). Speech recognition with support vector machines in a hybrid system. In *Proceedings of the Interspeech Conference* (pp. 993–996). Lisbon/Portugal.

Slaney, M. (1998). Auditory toolbox. *Apple Computer Technical Report, 45*.

Szepannek, G., Klefenz, F., & Weihs, C. (2005). Schallanalyse – Neuronale Repräsentation des Hörvorgangs als Basis. *Informatik Spektrum, 28*(5), 289–295.

Szepannek, G., Bischl, B., & Weihs, C. (2008). Towards automatic lyrics extraction from popular music. In C. Weihs, U. Ligges, A. Klapuri, and R. Martin, (organizers), *Int. Workshop on Music Signal Analysis*. (presentation) Witten, Nov. 3–4, 2008.

Szepannek, G., Harczos, T., Klefenz, F., & Weihs, C. (2009). Combining different auditory model based feature extraction principles for feature enrichment in automatic speech recognition. In *Proc. of the Specom 2009 Conference*. June 21–25. St. Petersburg. 205–210.

SVM Based Instrument and Timbre Classification

Sebastian Krey and Uwe Ligges

Abstract In this paper we propose a method that allows for instrument and timbre classification from a single tone. Features are derived from a pre-filtered time series divided into small windows. Afterwards, features from the (transformed) spectrum, Perceptive Linear Prediction (PLP), and Mel Frequency Cepstral Coefficients (MFCCs) as known from speech processing are selected. Clustering methods (e.g. k-means) are applied yielding a reduced number of aggregated features for the final classification task.

It turns out that a polynomial kernel with reasonable complexity can be used for the SVM. Accuracy of the results is very convincing given a misclassification error of roughly 19% for 59 different classes of instruments. Misclassification error is much smaller for a reasonable small number of classes, of course.

During methodological work, we ported the 'rastamat' library (Ellis 2005) functionality from Matlab to R. This means feature extraction as known from speech processing is now easily available from the statistical programming language R.

1 Introduction

A common task in Music Signal Processing is the recognition and classification of instrument tones. In this paper we propose a method for preprocessing and feature extraction to allow for a better classification of the recorded instruments than with common features.

Features are derived from a pre-filtered time series divided into small windows. Afterwards, feature selection is based on the (transformed) spectrum. Perceptive Linear Prediction Coding (PLP) (Hermansky 1990) and Mel Frequency Cepstral Coefficients (MFCCs) (Davis and Mermelstein 1980) are widely used in the context

S. Krey (✉)
Fakultät Statistik, Technische Universität Dortmund, 44221 Dortmund, Germany
e-mail: krey@statistik.tu-dortmund.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_84,

of speech recognition. Here they are used to model an instrument's tone. Other methods for instrument and timbre analysis have been described by Klapuri and Davy (2006).

In order to find a reduced number of aggregated features containing more specific information of tones, we aim at clustering windows of the (at least) 3 phases of a tone (attack, sustain, decay). Hence, clustering methods (e.g. k-means) are applied yielding aggregated features for the final classification task.

Using a Support Vector Machine (SVM) with polynomial kernel the classification gives convincing results. We used the implementation by Karatzoglou et al. in the R (R Development Core Team 2008) package 'kernlab' (Karatzoglou et al. 2004). On the single note recordings of the McGill Instrument Database (Opolko and Wapnick 1987) we achieve an misclassification error of 10% for classifying the instruments in 25 instrument families and 19% for discriminating between all 59 available instrument timbres in the database. Misclassification error is much smaller for a reasonable small number of classes, of course.

During methodological work, we ported the 'rastamat' library (Ellis 2005) functionality from Matlab to R. This means feature extraction as known from speech processing is now available from the statistical programming language R.

The used sounds from the McGill Instrument Database, which consists of 1986 notes (3–5 s long) played on 38 different instruments with different playing techniques (with or without vibrato, pizzicato or bowed, clean or distorted, etc.) resulting in 60 different timbres. For every timbre between 6 and 88 recorded notes are available, representing the tonal range of the instrument.

Based on this dataset two classification tasks are formed. One task is to discriminate between all instrument timbres. We drop the slapping and popping sounds of the electronic bass (only six examples available), resulting in a 59 class problem. For the other task the instruments are grouped in 25 instrument families (trumpets, flutes, bowed strings, etc.) resulting in an much easier classification problem.

2 Feature Extraction

To allow clustering and classification of sound recordings the extraction of feature vectors is necessary. We propose the usage of two approaches. First one is an autoregressive modelling approach called 'Perceptive Linear Prediction' (PLP), see Sect. 2.2. For the second approach we use 'Mel Frequency Cepstral Coefficients' (MFCC) that are derived from a spectral analysis of the transformed spectrum after preprocessing the data, see Sect. 2.3. All these features are known from speech processing and deliver promising results on recordings of musical instruments.

2.1 Preprocessing

The first preprocessing steps are identical for both sets of feature vectors.

Preemphasis Filtering

To compensate for the fact that the harmonics of a note produced by an instrument quite often have less energy than the fundamental frequency the sampled waveform of the recording is filtered using a digital one zero filter

$$y_t = x_t - 0.97x_{t-1}$$

resulting in a boost of higher frequencies of the original time series x_t.

Short Time Fourier Transformation

The second step is the transformation of the sampled waveform in the frequency domain using a Fourier Transformation. In order to keep some temporal information a Short Time- or windowed Fourier Transformation (STFT) is used:

$$F(t,k) = \sum_{j=1-M}^{N-M} w(j-t)y_j \exp\left(-2i\pi j\frac{k}{N}\right),$$

where N is the number of samples in the recording, t the time, k the number of the Fourier coefficient and $M = N - T$. To get smooth transitions at the borders of each window we apply the Hamming window function with a window width of T samples

$$w(t) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi t}{T}\right), & -\frac{T}{2} \le t \le \frac{T}{2} \\ 0 & \text{else} \end{cases}$$

to form the different frames of the STFT. The window width is a compromise between time and frequency resolution. We choose a width of 25 ms with 10 ms movement of the window in each step. For the further processing steps we drop the phase information and use the power spectrum.

Melscale Transformation

As a first information reduction step, the resulting Fourier frequencies of the STFT are transformed to the Melscale, a psychoacoustically motivated frequency scale, defined by

$$Mel(hz) = 2595\log_{10}\left(1 + \frac{hz}{700}\right),$$

and grouped into 40 equidistant frequency bins (from 0 to Nyquist frequency) on the Melscale.

2.2 Perceptive Linear Prediction

The construction of the Perceptive Linear Prediction (PLP) (Hermansky 1990) coefficients needs four further steps. First a loudness correction, resulting in equal loudness in all frequency bands, is performed through multiplying the amplitudes of the 40 frequency bins with the factor

$$L(f) = \left(\frac{f^2}{f^2 + 160\,000}\right)^2 \cdot \frac{f^2 + 1\,440\,000}{f^2 + 9\,610\,000}$$

where f is the center frequency of each bin.

Then a loudness compression through a cubic root transformation is performed. Afterwards each frame is transformed back into the time domain using an Inverse Fourier Transformation. The last step is to fit an autoregressive model, $y_t = \sum_{j=1}^{p} a_j y_{t-j} + e_t$, on every time frame. The resulting feature vector for each frame consists of the AR coefficients $(a_1, \ldots, a_p)$, where p defines the complexity of the model. We use $p = 8$ and $p = 15$.

2.3 Mel Frequency Cepstral Coefficients

The calculation of the Mel Frequency Cepstral Coefficients (MFCC) (Davis and Mermelstein 1980) consists of only two steps. First a loudness compression is realized through a logarithmic transformation of the 40 bins' amplitudes. Denote these grouped and transformed Fourier frequencies with z_t. Then a Discrete Cosine Transformation (DCT), $Z_k = \sum_{t=0}^{N-1} z_t \cos(\pi t\, k/N)$, of these transformed per frame spectra is calculated. The resulting feature vectors are the first q DCT coefficients $(Z_1, \ldots, Z_q)$. Here we use $q = 16$.

3 Clustering

After the preprocessing and feature extraction steps we have a p-dimensional feature vector for every frame of the Short Time Fourier Transformation. Motivated by the idea to model the three different phases *attack*, *sustain* and *decay* of an instrument's sound as well as in need of feature reduction, we apply clustering techniques to find representative feature vectors for every phase. We add an additional fourth cluster to handle the silence/noise parts that typically appear in the beginning/end of the recordings. In total we want to find the four mentioned clusters in every recorded note and use their cluster centers instead of the original per frame feature vectors, which greatly reduces complexity but still allows to make use of the change in the instruments sound. As the silence and noise cluster holds no useful information

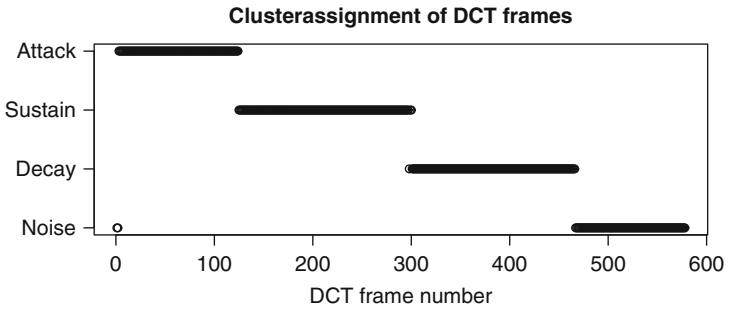


Fig. 1 Clustering of the DCT frames of a piano note

for the following classification task this cluster center is completely dropped for a further complexity reduction.

As clustering method we use k-means with 25 random start points, which gives us promising clustering results. In Fig. 1 the almost perfect clustering of a piano note is exemplarily shown.

4 Classification

The classification task is realised using a Support Vector Machine (SVM). A SVM searches for a separating hyperplane in the dataspace through solving the following optimization problem (Hsu et al. 2008)

$$\min_{w,b,\xi}\left(\frac{1}{2}w'w + C\sum_{i=1}^{n}\xi_i\right) \quad \text{subject to } \xi_i \geq 0,\ y_i(w'\varphi(x_i)+b) \geq 1-\xi_i$$

where $K(x_i, x_j) \equiv \varphi(x_i)'\varphi(x_j)$ is a kernel function and C the regularization parameter of the SVM. Using kernel functions, very complex nonlinear decision boundaries can be found. We have used the following four standard kernel functions in our search for the best working SVM for this classification task:

$$\begin{array}{ll}
\text{Linear} & K(x_i, x_j) = x_i' x_j \\
\text{Polynomial} & K(x_i, x_j) = (\sigma x_i' x_j + r)^d \\
\text{Gaussian RBF} & K(x_i, x_j) = \exp(-\sigma\|x_i - x_j\|^2) \\
\text{ANOVA RBF} & K(x_i, x_j) = (\sum_{k=1}^{p}\exp(-\sigma(x_{ik} - x_{jk})^2))^d
\end{array}$$

where $r, p \in \mathbb{R}$, $\sigma > 0$ and $d \in \mathbb{N}$.

The SVM classifier only supports binary classification tasks. To handle the multi-class problems in this work binary classifiers for all $\frac{n(n-1)}{2}$ combinations are trained and the final classifier is a majority vote of all these classifiers.

The estimation of the misclassification error is done in the outer run of a nested crossvalidation. In the inner run the best set of hyperparameters for the SVM is searched. Both crossvalidations are fivefold to reduce computation time.

5 Software

During this work we had to discover that for `R` (R Development Core Team 2008) only very few signal processing tools are available. The signal processing community developed their tools mainly for Matlab, resulting in quite a few implementations of the here used methods for feature extraction. As data transfer between different software packages is a time consuming task, we have ported the (for us) relevant parts of the function collection `rastamat` (Ellis 2005) to `R`, the software of our choice.

For the classification task we use the SVM implementation of the package `kernlab` (Karatzoglou et al. 2004). The classifiers' hyperparameter optimization and evaluation is performed with `mlr` (Bischl et al. 2009) a toolbox for easy performance comparison and optimization of different classification methods.

6 Results

Applying the presented methods on the recordings of the McGill instrument database delivers convincing results.

In Tables 1 and 2 the results of the 25 class problem, classifying the instruments into 25 instrument families are presented. As a reference the performance of a Linear Discriminant Analysis (LDA) and a Random Forest classifier are also listed. The Polynomial kernel function is the best performing kernel for the SVM in both cases, for PLPs and MFCCs. With a polynomial degree of $d = 3$ the PLPs achieve a misclassification error of 26%, which is comparable to the performance (Roever 2003) achieved in his work on the same classification problem.

Using the MFCCs the misclassification error can be reduced further. Here using a polynomial degree of $d = 2$ yields a misclassification error of 10%.

Table 1 Classification performance of PLPs on the 25 instrument families task

Classifier	Parameter	Error in %	Std.Dev. in %	Lag
SVM-Poly	$C = 1.4,\ d = 3$	26	3.7	8
SVM-RBF	$C = 1.4,\ \sigma = 0.133$	47	3.3	8
SVM-ARBF	$C = 1.4,\ \sigma = 0.142$	37	3.7	8
SVM-Lin	$C = 1.5$	38	2.9	8
RandFor	$U = 500,\ V = 7$	22	2.5	8
LDA		46	1.8	8

Table 2 Classification performance of MFCCs on the 25 instrument families task

Classifier	Parameter	Error in %	Std.Dev. in %
SVM-Poly	$C = 0.6,\ d = 2$	10	2.7
SVM-RBF	$C = 1.5,\ \sigma = 0.011$	11	2.7
SVM-ARBF	$C = 0.8,\ \sigma = 0.011$	10	2.2
SVM-Lin	$C = 1$	11	2.6
RandFor	$U = 1500,\ V = 7$	10	3.2
LDA		25	1.5

Table 3 Classification performance of PLPs on the 59 instrument timbres task

Classifier	Parameter	Error in %	Std.Dev. in %	Lag
SVM-Poly	$C = 1.6,\ d = 3$	43	2.6	8
SVM-RBF	$C = 1.5,\ \sigma = 0.122$	71	5.0	8
SVM-ARBF	$C = 1.4,\ \sigma = 0.117$	55	3.0	8
SVM-Lin	$C = 1.5$	51	2.8	8
RandFor	$U = 1500,\ V = 6$	34	3.9	8
LDA		60	2.5	8
SVM-Poly	$C = 1.4,\ d = 3$	47	2.6	15

Table 4 Classification performance of MFCCs on the 59 instrument timbres task

Classifier	Parameter	Error in %	Std.Dev. in %
SVM-Poly	$C = 1,\ d = 2$	19	3.0
SVM-RBF	$C = 1.5,\ \sigma = 0.011$	36	4.3
SVM-ARBF	$C = 0.6,\ \sigma = 0.11$	19	3.4
SVM-Lin	$C = 0.8$	20	3.2
RandFor	$U = 500,\ V = 7$	27	2.6
LDA		34	3.1

In both cases the Random Forest classifier is competitive to the SVM, whereas the LDA is lagging behind.

The classifier performance on the task to discriminate between all 59 instrument timbres are listed in Tables 3 and 4. The behaviour of the classifiers and the two different feature sets are quite similar to the 25 class task but with a naturally higher error rate. Here the PLPs achieve 43% and the MFCCs excellent 19% misclassification error.

The performance of the PLPs can be improved with a per frame standardization of the autoregressive coefficients a_i before the clustering step. With this modification the SVM with Polynomial Kernel reaches 33% misclassification error. The Random Forest cannot benefit from this modification.

Increasing the Lag of the autoregressive model of the PLPs to $p = 15$ or combining MFCCs and PLPs does not result in lower misclassification rates.

The best methods achieve a very good classification performance comparible to other methods or human listeners as summarized in Klapuri and Davy (2006).

7 Conclusion

In this paper we have presented a method to use features from speech processing for instrument and timbre classification. These features are based on the windowed and transformed spectrum of the input signal.

Through clustering we could reduce the original per window feature vectors drastically to three cluster centers, which represent the different phases *attack*, *sustain* and *decay* of an instruments sound. With these cluster centers as input variables we could use a Support Vector Machine with polynomial kernel to classify the single note recordings of the McGill instrument database. On the 25 classes problem to discriminate between the different instrument families, we achieved a misclassification error of 10%. The task to discriminate between all in the McGill database available 59 different instrument timbres resulted in a misclassification error of 19%. All these results were estimated using a nested crossvalidation, avoiding the usage of the same partition of the dataset for optimization of the SVM hyperparameters as well as error estimation, which could give overly optimistic errorrates.

During this work we have ported the necessary parts of the Matlab function collection `rastamat` to `R`, giving us the possibility to realize the whole numerical calculations without the time consuming need to transfer data between different software packages.

References

Bischl, B., Wornowizki, M., & Borg, K. (2009). *The mlr Package: Machine Learning in R*. From http://www.algorithm-forge.com/bischl/mlr/.

Davis, K., & Mermelstein, P. (1980). Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics Speech and Signal Processing, ASSP-28*(4), 357–366.

Ellis, D. P. W. (2005). *PLP and RASTA (and MFCC, and inversion) in Matlab*. From http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal of Acoustical Society of America, 87*(4), 1738–1752.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2008). *A practical guide to support vector classification*. Taipei: National Taiwan University. From http://www.csie.ntu.edu.tw/~cjlin.

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab – An S4 package for kernel methods in R. *Journal of Statistical Software, 11*(9), 1–20. From http://www.jstatsoft.org/v11/i09/.

Klapuri, A., & Davy, M. (2006). *Signal processing methods for music transcription*. New York: Springer.

Opolko, F., & Wapnick, J. (1987). *McGill University master samples (CDs)*. Quebec, Canada: McGill University.

R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0, from http://www.R-project.org.

Roever, C. (2003). *Musikinstrumentenerkennung mit Hilfe der Hough-Transformation*, Diplomarbeit TU Dortmund, Fakultät Statistik. From http://www.aei.mpg.de/~chroev/publications/RoeverDiplom.pdf.

Three-way Scaling and Clustering Approach to Musical Structural Analysis

Mitsuhiro Tsuji, Toshio Shimokawa, and Akinori Okada

Abstract In the present paper, we propose and investigate three-way scaling and clustering approaches to musical structural analysis of emotion expression. The approach is applied to the affective values of music by subjects who were interested in theater and in instrumental music as well. We expected that we would find interesting structure in the affective expression of music from three view points: affective values, individuals, and music. We analyzed the three-way structure model by applying INDSCAL and INDCLUS. We expected that the INDSCAL model would show a geometrical structure which offers interesting insights about the characteristics of affective values. Furthermore we expected that the INDCLUS model would reveal further geometrical structure.

1 Introduction

The field of statistics in music research is very broad (Weihs et al. 2007). The data we were going to deal with were responses by 14 subjects along 24 affective values on 30 instrumental music pieces (Table 1). The 14 subjects were interested in theater and were thought likely to give high affective values (Fig. 1) (Kishihara and Tsuji 2006).

First, we explain the characteristic interpretation which shows the structure of affective judgements. We use factor analysis as an exploratory approach to compare the result with that of INDSCAL. We then explain the geometrical interpretation of the results of INDSCAL from the view point of various factors. Finally, we will validate some additional interpretations of the results of INDCLUS from the view point of the classification of the real structure. Of course, factor analysis and INDSCAL are different, but from a psychological view, we can obtain a geometrical interpretation of both. Taniguchi (1998) applied factor analysis to obtain structures from

M. Tsuji (✉)
Kansai University, Osaka, Japan
e-mail: tsuji@kansai-u.ac.jp

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_85,

Table 1 Musical Stimuli: instrumental music pieces under discussion

No.	Music	Time	No.	Music	Time
1	The First Noel	4:20	16	Exterminate	2:55
2	CUMBA's DANCE	2:02	17	Beautiful	4:30
3	ATLANTIC WEAVE	6:05	18	Missing Link	2:22
4	Zontac Hill	5:00	19	The Dew of Life	3:35
5	Going Under	3:31	20	Calling You	4:38
6	Scarbough Fair	4:11	21	Eternal Prayer	5:12
7	Another Star	3:50	22	Sleepless Beauty	2:49
8	Author of Life Divine	2:27	23	Scarborough Fair (Akasa)	3:37
9	Temptation	2:00	24	FOR FOURTY DAYS	5:30
10	Solitude	4:44	25	Their Daily Lives	4:09
11	Love Letter to Andes	4:11	26	TNR	1:53
12	The Fool On The Hill	2:50	27	Samurai Faith	4:53
13	Follow Me Up To Carlow	3:57	28	Guilty	3:43
14	Beat of Dream	4:55	29	Scarborough Fair (Orig)	4:00
15	Scarborough Fair (Sala)	4:08	30	Accustom	1:15

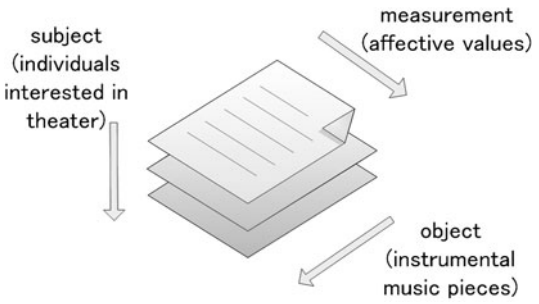


Fig. 1 Three-way data under discussion

affective judgements. He dealt with the data of 24 affective values in music collected from 183 students (subjects) on five classical music pieces. The resulting five factors were " Enhancement", "Indiscretion", "Intensity", "Solemnization" and "Affinity". We applied the same method of Taniguchi (1998) to our data (Fig. 2).

We found several characteristic interpretations among the affective judgement;

- Five factors exist.
- Four variables (Cheerful, Joyous, Happy, and Bright) belong to "Enhancement (+)" or "Indiscretion".
- One variable, Restless, may belong to "Indiscretion", "Intensity" or "Solemnization".
- One variable, Calm, may belong to "Affinity" or "Intensity".

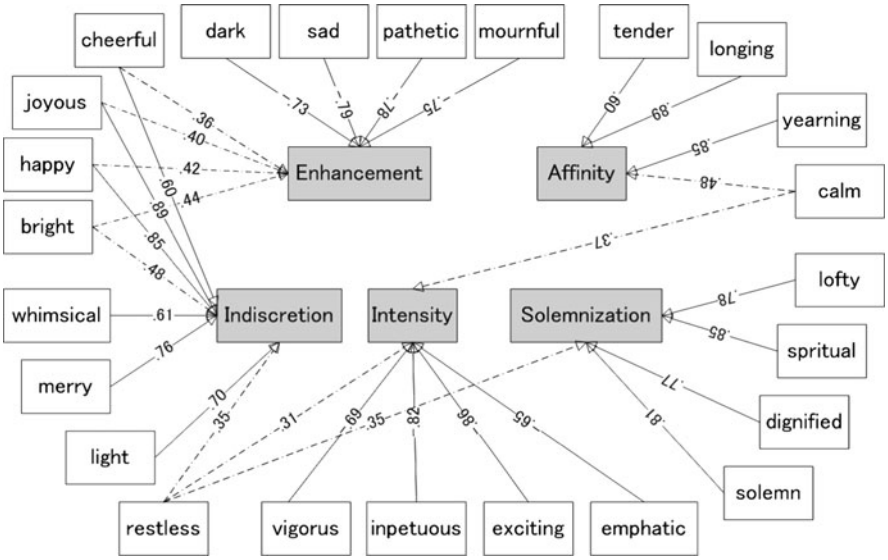


Fig. 2 Results (factor loading) using a traditional factor-analysis model

2 The Three-Way Structure Model

2.1 Results by INDSCAL

As the first step, we expected that INDSCAL might show a geometrical structure which would offer some interesting insights concerning the affect of music. In the INDSCAL model, the three-way matrix $D \equiv \{d_{ij,k}\}$ is estimated from an input data matrix $\Delta \equiv \{\delta_{ij,k}\}$, where $d_{ij,k}$ is the distance between affective values i and j for individual k in the common stimulus space, and $\delta_{ij,k}$ is the dissimilarity between affective values i and j for individual k. Distance $d_{ij,k}$ is represented by $d_{ij,k} = \sqrt{\sum_{r=1}^{R} w_{kr}(x_{ir} - x_{jr})^2}$, where w_{kr} is the weight of importance along dimension $r (r = 1, \ldots, R)$ for individual $k (k = 1, \ldots)$, x_{ir} and x_{jr} are coordinates of the affective values i and j along dimension r of an R-dimensional common stimulus space.

Using INDSCAL (Arabie et al. 1987), more than 1,000 initial coordinate values served as input. Output diagrams are shown in Figs. 3 and 4. In comparison to the result of factor analysis, we found certain structures in the affective values, as follows;

- It seems that there are six factors.
- The variable, Restless, is located at a strange position.
- The variable, Calm, seems to belong to “Affinity”.

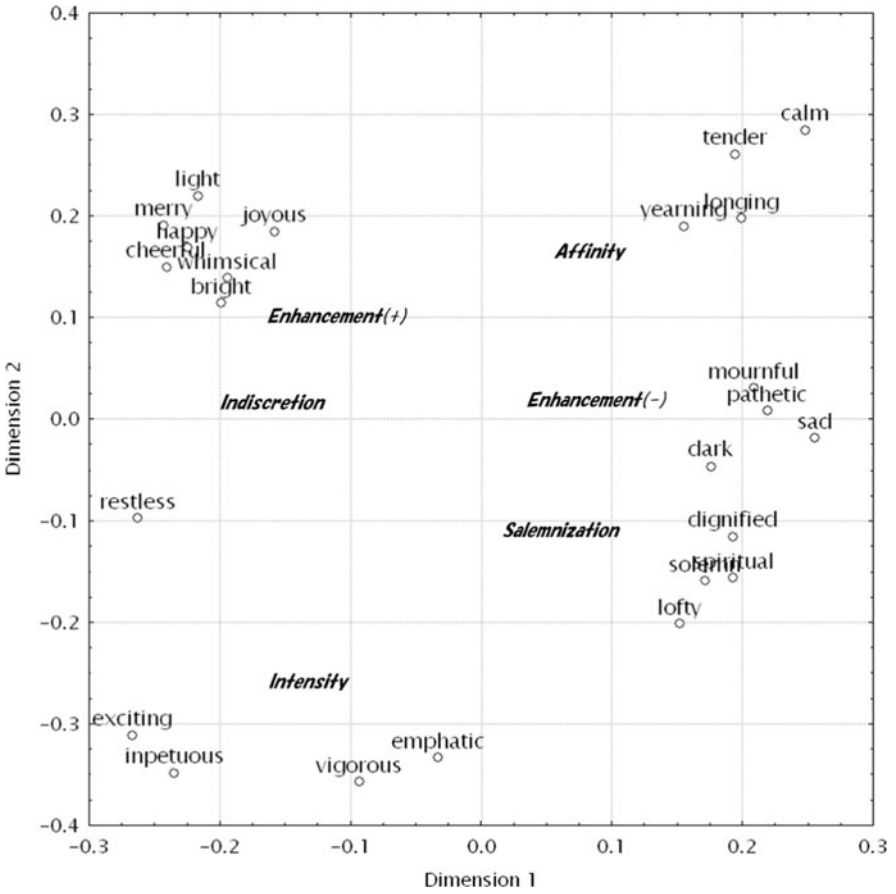


Fig. 3 Two-dimensional plot of affective values by INDSCAL (VAF = 0.701)

- In two dimensional projections of three-dimensional solution (lower part of Fig. 4), we find that the bidirectional two factors ("Enhancement (+)" and "Enhancement (−)") seem to merge into one cluster.

2.2 Results Using INDCLUS Presented as a Hanabi Chart

We expected that INDCLUS might reveal a detailed geometrical interpretation of the three-way structure (Fig. 1). The INDCLUS model is represented by $s_{ij,k} \cong \sum_{r=1}^{R} w_{kr} p_{ir} p_{jr} + c_k$, where $s_{ij,k}$ is the similarity between affective values i and j for individual k ($k = 1, \ldots, K$); w_{kr} is the weight of individual k for cluster r, p_{ir} represents whether affective value i beings to cluster r ($p_{ir} = 1$) or not ($p_{ir} = 0$), and c_k is an additive constant of individual k. The characteristics of the INDCLUS

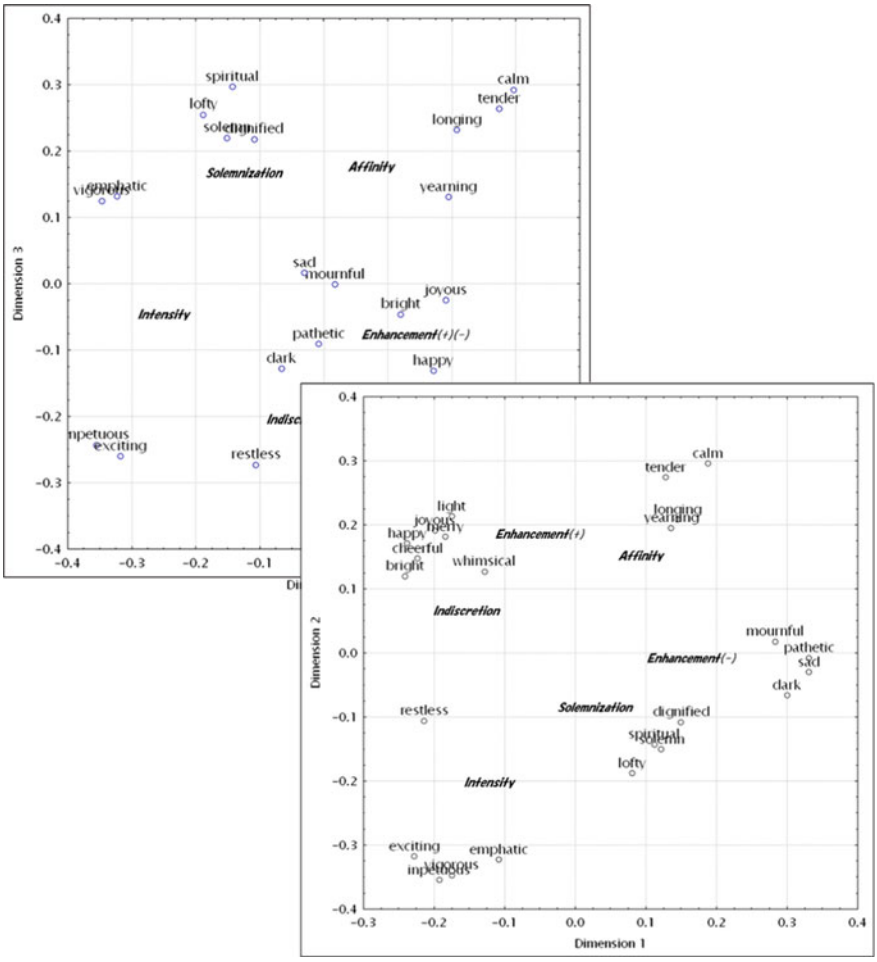


Fig. 4 Three-dimensional plot of affective values using INDSCAL (VAF = 0.768)

model are overlapping clustering and non-hierarchy. The analysis using INDCLUS (Arabie et al. 1987) was done with more than 500 initial values served as input. The integrated approach of three-way analysis (INDSCAL and INDCLUS) is an effective tool for exploratory analysis, and reveals interesting structures in affective values as follows:

- Strong connection among variables which do not belong to any other cluster.
- Outlier variables which do not belong to any specific cluster.

For the agglomerative hierarchical clustering method, the contour line figure reveals a simultaneous presentation of structures. For the INDCLUS model Hanabi (fire works) Chart simultaneously presents the INDSCAL and INDCLUS results. Hanabi Chart consists of two or more figures corresponding to the number of

clusters. In each chart, we make a bridge by a dotted line between cluster number and affective variable which belongs to the specific cluster. This cluster number is positioned on extension of the line from the origin to the mean affective value. We named Hanabi Chart because the bridge by dotted lines resemble the light of fire works.

In Fig. 5, we show the typical structure in the two-dimensional solutions of INDSCAL which we can see in the INDCLUS analysis including from two through four clusters.

Variables Which do not Belong to any Cluster at First

When the number of clusters is small, there are some variables which do not belong to any cluster. We put our attentions to four variables (Vigorous, Impetuous, Exciting, and Emphatic), at the bottom of the uppermost panel of the Hanabi Chart (Fig. 5) which are not connected to any cluster through dashed lines.

When the number of clusters is not so small, these four variables are connected to the same cluster at least once. So we can confirm that these four variables belong to the same cluster.

Variable Which does not Belong to any Specific Cluster

Often some variables may not belong to any specific cluster. In our data, variable, Restless, does not belong to the original group "Indiscretion". We find the existence of this strange variable in Fig. 5. In the case of two clusters, variable Restless is connected to the same cluster C-2 as three other variables, Whimsical and Merry and Light, which belong to group ("Indiscretion"). In the case of four clusters, variable Restless is not connected to the same cluster.

3 Conclusion and Discussion

We utilized a three-way scaling and clustering approach to investigate musical structural analysis of emotion expression.

The approach was applied to the affective values in music derived by subjects interested in theater and instrumental music. We have analyzed this three-way structure model by applying INDSCAL and INDCLUS, and have found that INDSCAL shows a geometrical interpretation which offers some interesting insights about the subjects' affective judgement. We found that INDCLUS shows the detailed geometrical interpretation of this structure. Moreover we have introduced an Hanabi Chart which supports simultaneous presentation of INDSCAL and INDCLUS results.

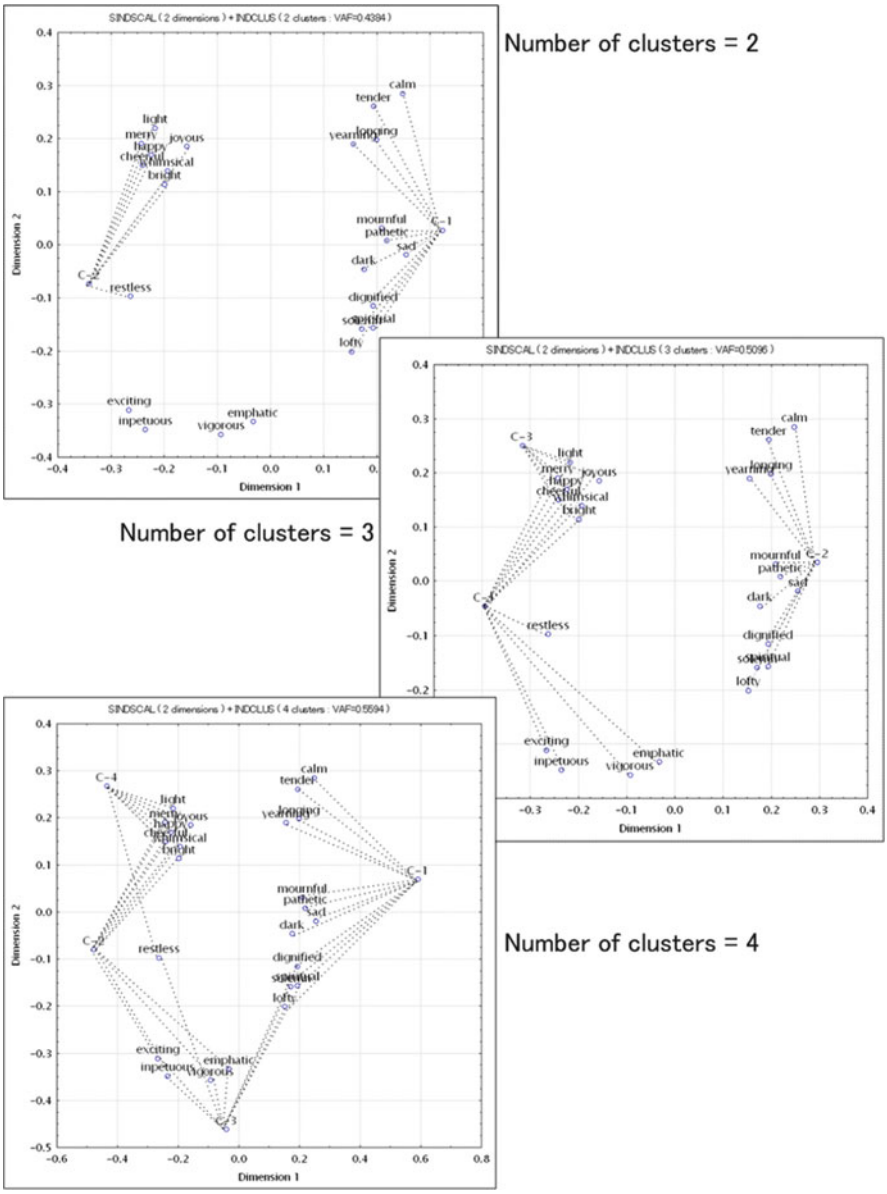


Fig. 5 Affective value two-dimensional and two-four clusters plot using INDSCAL and INDCLUS (VAF = 0.438, 0.510, 0.537)

We found a variable, Restless, which does not belong to any specific cluster. By removing this variable in the next analysis (Fig. 6), we could improve the VAF value (goodness of fit) from 0.608 into 0.695.

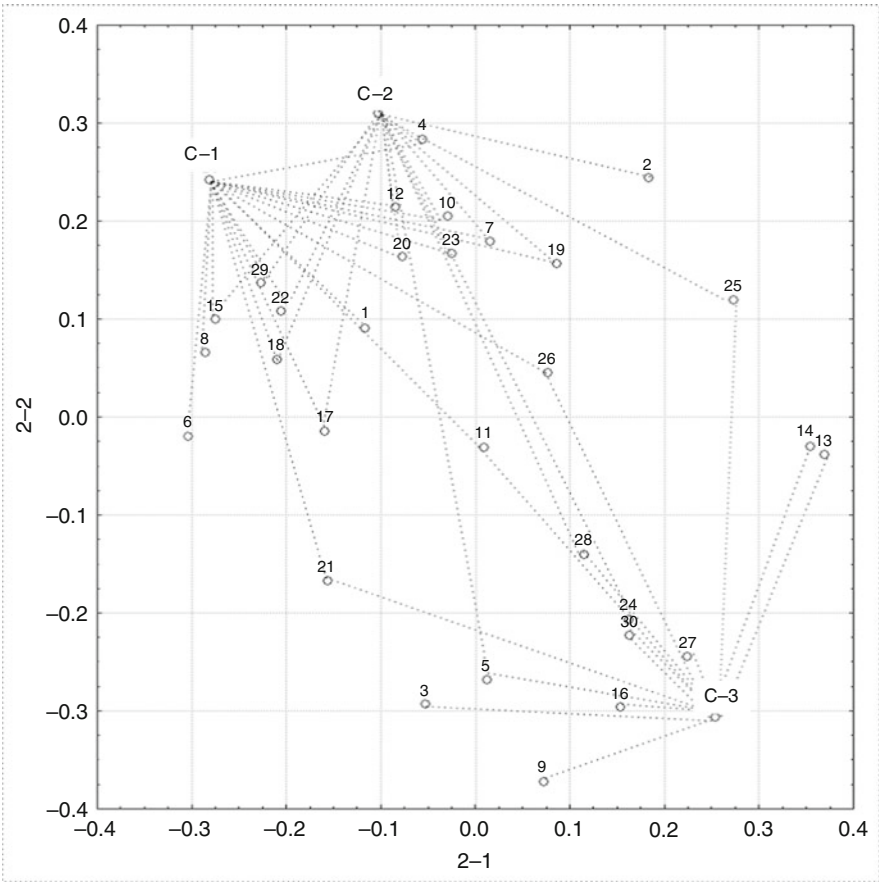


Fig. 6 Music two-dimensional and three-cluster plot using INDSCAL (VAF = 0.695) and INDCLUS

Acknowledgements This work was supported by a "Strategic Project to Support the Formation of Research Bases at Private Universities": Matching Fund Subsidy from MEXT (Ministry of Education, Culture, Sports, Science and Technology), 2008–2012.

References

Arabie, P., Carroll, J. D., & DeSarbo, W. S. (1987). *Three-way scaling and clustering*. Newbury Park: Sage.

Kishihara, M., & Tsuji, M. (2006). A classification trial of stage music by affective value. *MUS-67*, 33–36 (in Japanese).

Taniguchi, T. (1998). *Music and emotion*. Kyoto: Kitaouji (in Japanese).

Weihs, C., Ligges, U., Morchen, F., & Mullensiefen, D. (2007). Classification in music research. *Advances in Data Analysis and Classification, 1*(3), 255–291.

Improving GMM Classifiers by Preliminary One-class SVM Outlier Detection: Application to Automatic Music Mood Estimation

Hanna Lukashevich and Christian Dittmar

Abstract Automatic estimation of music mood has emerged as an important task in Music Information Retrieval. It has direct applications in music search engines and cross-modal multimedia tools. During the last years, Gaussian Mixture Models (GMM) became one of the most popular classifiers for mood estimation. One of the remaining key challenges is the impossibility to collect representative training data sets. With GMM classifiers, "unknown" test data can result in low log-likelihoods for all mood classes, so that the resulting decision becomes immethodical. Thus, we suggest using a preliminary outlier detection based on one-class Support Vector Machines (SVM). In this paper we introduce a novel approach to optimize the one-class SVM parameters via minimizing the differences between the fraction of outliers, fraction of support vectors and parameter ν.

1 Introduction

During recent years the scientific and commercial interest in Music Information Retrieval (MIR) has significantly increased. Stimulated by the ever-growing availability and size of digital music collections, automatic music mood estimation has been identified as an increasingly important means to aid convenient exploration of large music catalogs. Online music shops and content aggregators have realized that search functionality beyond conventional metadata such as artist, title and album is a very effective tool to push unknown or long-tail (Celma 2008) content. Especially for professional applications, such as TV post production and program planning in radio stations, automatically derived mood tags can significantly enhance organization and accessibility of content. Aupeo[1] and Musicovery[2]

[1] http://www.aupeo.com
[2] http://www.musicovery.com

H. Lukashevich (✉)
Fraunhofer IDMT, Ehrenbergstr. 31, 98693 Ilmenau, Germany
e-mail: lkh@idmt.fraunhofer.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_86,

are examples of innovative services that allow end-users to listen to personalized mood based Internet radio streams.

1.1 Mood Models

Many publications have addressed suitable modeling methods for musical mood, although it is obvious that the human perception of music mood as a subjective, context dependent and multi-dimensional concept that can not be modeled to the utmost extent. Generally, moods are not well-defined and cannot be unambiguously identified by human listeners. Mood models reported in the literature can be roughly divided into category-based, dimension-based and combinations of both. Early work on music expression concentrates on category based formalization, such as Hevner's adjective circle (Hevner 1936). Category based approaches allow the assignment of music items into one or multiple groups which results in a single- or multi-label classification problem. Dimension-based mood models focus on the description of mood as a point within a multi-dimensional mood space, commonly using dimensions such as valence and arousal. As an example, Thayer's model (Thayer 1989) uses the dimensions energy and tension. Mood models, that combine categories and dimensions typically place mood adjectives in a region of the mood space, e.g. the Tellegen-Watson-Clark model (Tellegen et al. 1999).

1.2 Mood Audio Features

Publications on automatic mood classification report a variety of acoustic features. Lu et al. (2006) utilize various rhythmic features such as Rhythm Strength, Average Tempo and Average Onset frequency. Li and Ogihara (2004) and Tolos et al. (2005) use spectral features (e.g. Mel-Frequency Cepstral Coefficients (MFCC), Audio Spectrum Centroid (ASC) and others) to describe the timbre. Furthermore, Wu and Jeng (2008) setup a complex mixture of a wide range of acoustical features for valence and arousal estimation: Rhythmic Content, Pitch Content, and others. In the system described in this paper, the following features are used: Log Loudness, Norm Loudness, MFCC, Audio Spectrum Envelope, ASC, Spectral Crest Factor, Spectral Flatness Measure and Zero Crossing Rate. Besides this set of low-level features, several mid-level representations (Dittmar et al. 2007) are used. These mid-level features are specialized in the domains timbre, rhythm and tonality. They range from simple modulation coefficients, to auto-correlation-based rhythmic patterns, to histograms of note and chord candidates derived from a chromagram, based on Enhanced Pitch Class Profiles (Lee 2006).

1.3 Mood Classification

Well described machine learning algorithms such as GMM and SVM are most commonly used in the literature. The discriminative SVM approach is used in Li and Ogihara (2004). Trohidis et al. (2008) compare different multi-label classification schemes based on an SVM and k-Nearest Neighbor classifier. Examples for GMM based approaches are given in Lu et al. (2006) and Zhang et al. (2003). In Dunker et al. (2008), GMM and SVM classifiers are compared with a slightly better result for the SVM approach. A fundamental issue with conventional GMM-based single-label classification is the assumption that every observed data sample is generated by one, and only one class. Thus, it is impossible to assemble enough training data for this Open-World (the number of classes is unbound) classification problem in practice. In fact, the Open-World problem is mostly simply treated as Closed-World (the number of classes is fixed) problem. Although this simplification is working for a number of classification tasks, it is by no means justified. As will be detailed in the following sections, we attempt to turn the Open-World problem into a real Closed-World problem by constraining the region of interest.

2 Proposed System

As mentioned above, it is not feasible to collect sufficient representative training data sets for GMM-based mood classification. In topological terms, undefined areas remain in the feature space. If test data exhibits properties that the model has not been trained with it can result in low log-likelihoods for all mood classes in parallel. Thus, the resulting decision which class to favor becomes immethodical. To tackle this problem we propose an outlier detection algorithm based on one-class SVM. Basically, we can deduce some information about our region of interest, which is defined as the partition of the feature space, where the majority of training data is situated. The work flow of the proposed system is depicted in Fig. 1. On the training stage we extract acoustical feature vectors as described in Sect. 1.2. We apply a feature space transformation (FST) to reduce the dimensionality of the feature space. In this work, Linear Discriminant Analysis (LDA) (Webb 2002) is used. The transformation matrix of LDA obtained on the training stage is re-used in the testing stage to perform the FST. Each mood class is modeled with a multivariate GMM. In parallel we train a one-class SVM on the entirety of feature vectors (i.e., across all classes) to obtain the hyper region of the feature space containing the majority of training data (see Sect. 3 for details). In the testing stage we use one-class SVM to classify the feature vectors in so-called *targets* (data points belonging to class "1" of the one-class SVM) and *outliers* (data points belonging to class "−1" of the one-class SVM). Consequently, we apply GMM to classify all test data, targets and outliers and achieve three accuracy values. Comparison of these accuracy values allows us to assess if one-class SVM indeed rejects outliers.

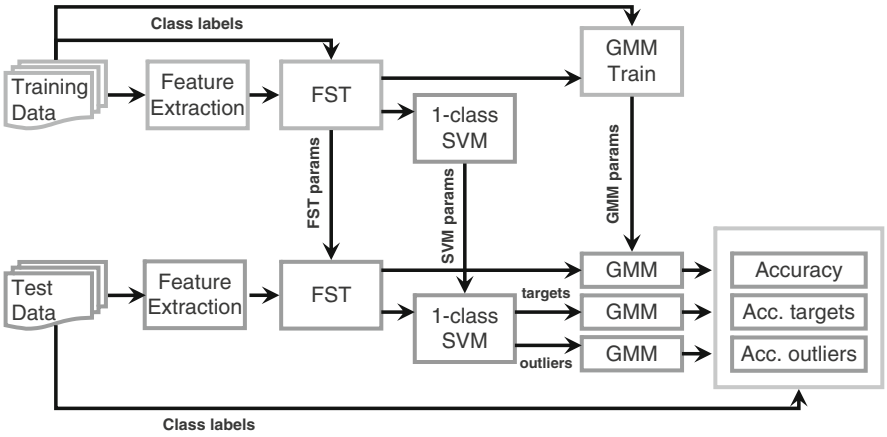


Fig. 1 Work-flow of evaluation framework

3 Outlier Detection with One-Class SVM

3.1 One-Class SVM

One-class SVM was firstly proposed by Schölkopf et al. (2001) for estimating the support of a high-dimensional distribution. As in the case of a two-class SVM, the kernel function is used to map the feature vectors into a higher dimensional space. By utilizing one-class SVM, most of the data are separated from the origin by a large margin in the higher dimensional space. Given the training vectors $\mathbf{x}_i \in R^n$, $i \in [l]$, the model is estimated in the following way:

$$\min_{\mathbf{w},b,,\rho} \frac{1}{2}\mathbf{w}^T\mathbf{w} - \rho + \frac{1}{\nu l}\sum_{i=1}^{l}\xi_i \qquad (1)$$
$$\text{subject to } \mathbf{w}^T\phi(\mathbf{x}_i) \geq \rho - \xi_i,\ \xi_i \geq 0,$$

where $\rho/\|\mathbf{w}\|$ specifies the distance from the decision hyperplane to the origin, and ξ_i are introduced slack variables. The trade-off parameter $\nu \in (0, 1]$ corresponds to an expected fraction of outliers within the feature vectors. A solution of the system (1) enables the usage of the decision function: $\text{sgn}\left(\sum_{i=1}^{l}\alpha_i K(\mathbf{x}_i, \mathbf{x}) - \rho\right)$, where $K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ is a kernel function, equivalent to a dot product of the feature vectors mapped into the higher dimensional space; and $\boldsymbol{\alpha}$ is a vector with Lagrange multipliers, needed to solve (1). In the experiments in this work, we use the most common type of kernel, namely Radial Basis Function (RBF): $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$

3.2 Estimation of Kernel Parameters

We propose a novel approach to optimize one-class SVM kernel parameters. Due to the lack of class labels in the one-class SVM, it is not possible to optimize the kernel parameters using cross-validation, as it is common for a two-class SVM. In the case of the RBF kernel, we have to tune the kernel parameter γ. Different methods have been reported for choosing the best parameters for an one-class SVM. In Tax and Duin (2001) an estimate for the volume covered by the one-class classifier is obtained. In Tran et al. (2005) the fraction of support vectors is chosen to characterize the precision of one-class SVM. Xie (2006) utilized a weighted sum of two kind of errors, namely fraction of support vectors and fraction of outliers. It was shown by Schölkopf et al. (2001), that the trade-off parameter ν is an upper bound on the fraction of outliers and a lower bound to the fraction of support vectors. Our optimization method is based on minimizing the differences between the fraction of outliers f_{out}, the fraction of support vectors f_{SV} and the trade-off parameter ν. The optimal γ minimizes $F = \lambda \cdot (f_{SV} - \nu) + (1 - \lambda) \cdot (f_{out} - \nu)$, where $\lambda \in [0, 1]$. Figure 2 demonstrates how the selection of the kernel parameter γ influences the resulting decision hyperplane of one-class SVM. Panel (a) shows the decision hyperplanes for three γ values. The objective function F, f_{SV}, f_{out} and volume fraction f_{vol} for different values of γ are represented in panel (b). Our observations show, that γ values corresponding to the minimal F reside on the left side of the "valley" of the objective function, while the optimal working point lies on the right side of the "valley". To account for this property, we use an additional tolerance parameter, shifting the working point to the right until the maximum allowed loss of accuracy is achieved. In this work, the additional accuracy loss is set to 0.01.

4 Evaluation

4.1 Dataset and Parameter Settings

We used a music set compiled by expert listeners. It comprises the following 10-fold mood taxonomy: Aggressive, Calm, Stressful, Danceable, Dramatic, Energetic, Happy, Melancholic, Fun and Relaxing. In total, the original feature space spanned by concatenating the features described in Sect. 1.2 is $d = 167$ dimensional. The total number of data samples is 4361 thus, around 430 data vectors represent each class. We simulate the lack of training data by restricting the fraction of the data set used for training. This fraction of data taken for training the models (further notated as p) is varied within $p = \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$. For example, $p = 0.3$ corresponds to using randomly chosen 30% of the data set for training the models and testing with the remaining 70%. We use LDA to reduce d to 9 dimensions. We varied the number of mixtures used for the GMM according to $m = \{3, 5, 10, 15, 20\}$, wheras optimal results with our test scenario were achieved for $m = 15$. The best

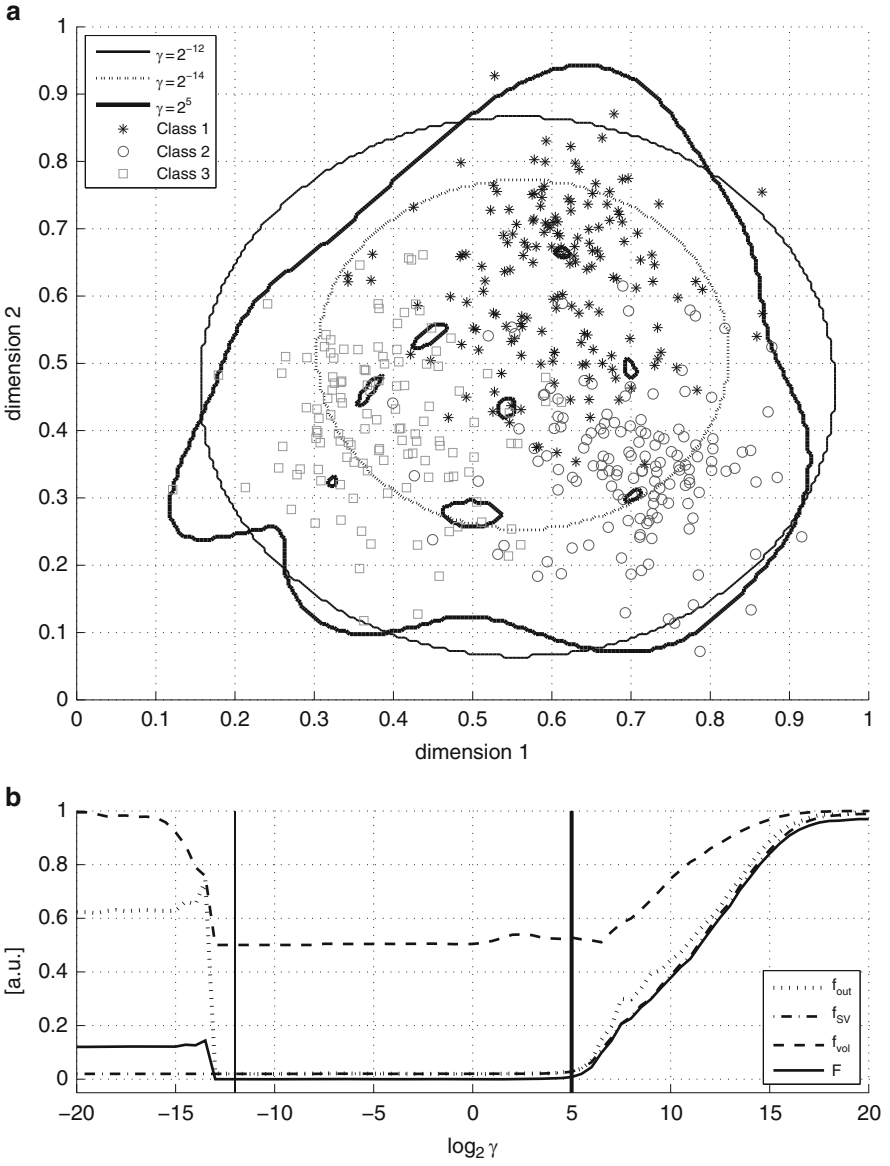


Fig. 2 Selection of the kernel parameter γ and its influence on the decision hyperplane of one-class SVM. Panel (**a**) Exemplary decision hyperplanes for three γ values in an LDA-transformed, two dimensional feature space. The hyperplane corresponding to $\gamma = 2^5$ is considered to be optimal. Panel (**b**) Fraction of support vectors, fraction of outliers, volume fraction and decision function F for different values of kernel parameter γ. *Vertical lines* mark γ values corresponding to the hyperplanes depicted in panel (a). Trade-off parameter ν is set to 0.02, weighting parameter λ for F calculation is set to 0.7

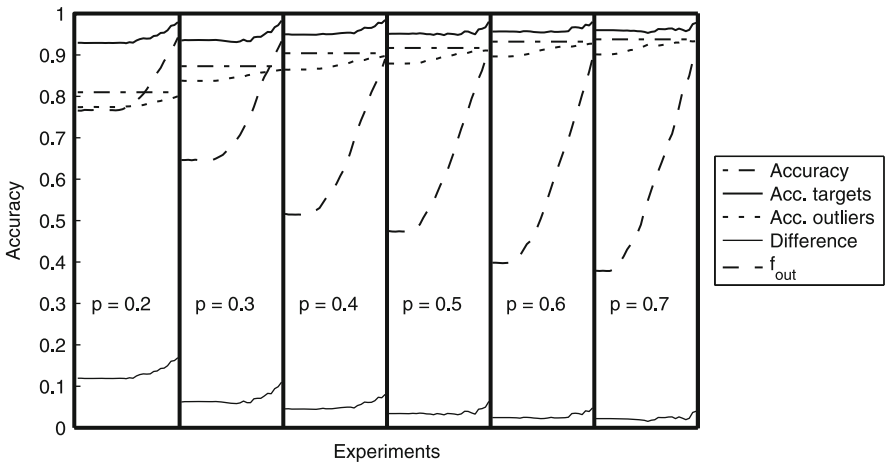


Fig. 3 Accuracy results for the proposed approach. Vertical blocks correspond to the results corresponding to various p values. Within each p-block, the trade-off parameter ν is varied from left to right within a range of $0.01 - 0.90$. For $p = 0.2$, the highest increase in accuracy is achievable

results for our data set have been achieved with a kernel parameter $\gamma = 0.25$. The last parameter is the expected number of outliers ν, it is varied in the range $\{0.01, \ldots, 0.90\}$.

4.2 Results

As can be seen in Fig. 3, the effect of the proposed method is most pronounced in case of low p, i.e., a small training set. Thus, it is applicable for real world classification problems. As expected, the small training set decimates the generalization capabilities of the one-class SVM, as can be seen from the increasing elevation of the total accuracy. The trade-off for improved target accuracy is a large fraction of outliers f_{out}, i.e., data samples that are rejected by the one-class SVM. Thus, the method is applicable for the tasks where a recall is not crucial, e.g., mood estimation scenarios. As an example, for $p = 0.5$, approximately 5% increase of target accuracy is achievable when tolerating 50% rejected data.

5 Conclusions

In this publication, an outlier detection algorithm based on one-class SVM has been described as pre-processing for GMM classification. This novel approach to optimization of one-class SVM parameters shows promising results. The method is useful in case of small training sets and for classification tasks, where low recall is tolarable.

Acknowledgements This work has been partly supported by grant No. 01MQ07017 of the German THESEUS program, funded by the Federal Ministry of Economics and Technology.

References

Celma, O. (2008). *Music recommendation and discovery in the long tail*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain.

Dittmar, C., Bastuck, C., & Gruhne, M. (2007). Novel mid-level audio features for music similarity. In *Proceeedings of the International Conference on Music Communication Science (ICOMCS)*, Sydney, Australia (pp. 38–41).

Dunker, P., Nowak, S., Begau, A., & Lanz, C. (2008). Content-based mood classification for photos and music: A generic multi-modal classification framework and evaluation approach. In *Proceedings of the International Conference on Multimedia Information Retrieval (ACM MIR)*, Vancouver, Canada (pp. 97–104).

Hevner, K.(1936). Experimental studies of the elements of expression in music. *American Journal of Psychology, 48*(2), 246–268.

Lee, K. (2006). Automatic chord recognition from audio using enhanced pitch class profile. In *Proceedings of the International Computer Music Conference (ICMC)*, New Orleans, USA.

Li, T., & Ogihara, M. (2004). Content-based music similarity search and emotion detection. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, *5*, 705–708.

Lu, L., Liu, D., & Zhang, H. (2006). Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech & Language Processing, 14*(1), 5–18.

Tax, D. M. J., & Duin, R. P. W. (2001). Uniform object generation for optimizing one-class classifiers. *Journal of Machine Learning Research, 2*, 155–173.

Tellegen, A., Watson, D., & Clark, L. A. (1999). On the dimensional and hierarchical structure of affect. *Psychological Science, 10*, 297–303.

Thayer, R. E. (1989). *The biopsychology of mood and arousal*. Oxford: Oxford University Press.

Tolos, M., Tato, R., & Kemp, T. (2005). Mood-based navigation through large collections of musical data. In *Proceedings of the 2nd IEEE Consumer Communications and Networking Conference*, Las Vegas, Nevada, USA (pp. 71–75).

Tran, Q. A., Li, X., & Duan, H. (2005). Efficient performance estimate for one-class support vector machine. *Pattern Recognition Letters, 26*, 1174–1182.

Trohidis, K., Tsoumakas, G., Kalliris, G., & Vlahavas, I. (2008). Multilabel classification of music into emotions. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*, Philadelphia, Pennsylvania, USA (pp. 325–330).

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation, 13*(7), 1443–1471.

Webb, A. R. (2002). *Statistical pattern recognition* (2nd ed.). London: Wiley.

Wu, T. L., & Jeng, S. K. (2008). Probabilistic estimation of a novel music emotion model. In *14th International Multimedia Modeling Conference*. Berlin: Springer.

Xie, L. (2006). Swarm intelligent tuning of one-class ν-SVM parameters. In *Proceedings of the 1st International Conference on Rough Sets and Knowledge Technology (RSKT)* (pp. 552–559).

Zhang, H. J., Liu, D., & Lu, L. (2003). Automatic mood detection from acoustic music data. In *Proceedings of the International Symposium Music Information Retrieval (ISMIR)* (pp. 81–87).

Multiobjective Optimization for Decision Support in Automated 2.5D System-in-Package Electronics Design

Martin Berger, Michael Schröder, and Karl-Heinz Küfer

Abstract We propose a multiobjective optimization approach for decision support in the 2.5D System-in-Package (SiP) design automation. 2.5D SiP is a relatively new integration concept for miniaturization in which a microelectronic system is integrated on vertically stacked substrate modules. We approach the SiP layout process with Pareto optimization. A database of optimized SiP layouts is interactively explored with our decision support tool 3D SiP Expert. Our optimization methods streamline the layout process, eliminate time-consuming redesign steps and support selecting SiP technologies and the ideal number of substrate modules.

We discuss a constructive heuristic that places the devices on the substrates. Our computational results show that the heuristic is efficient and finds solutions within 6% of optimality. We also propose group constraints that implicitly cluster device groups, e.g., functional cooperating devices, and that structure the placement.

1 Introduction

2.5D System-in-Package (SiP) is a relatively new integration concept for miniaturization in which discrete devices (e.g., passives, diodes, antennas) and integrated circuits of an electronic system are placed on vertically stacked substrate modules. SiP aims for cost-effective and flexible developed microsystems demanded for mobile communication, medical applications and the computer memory market. Figure 1 exemplarily shows two SiP technologies.

While engineers apply highly developed software tools for other integrations, there is a lack of tools for the design of SiPs (Polityko 2008). Deficiencies arise in the three stages of the layout process: the partitioning, the placement, and the routing. Figure 2 illustrates the layout stages in SiP design.

M. Berger (✉)
Fraunhofer Institute for Industrial Mathematics (ITWM), Fraunhofer-Platz 1,
67663 Kaiserslautern, Germany
e-mail: martin.berger@itwm.fraunhofer.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_87,

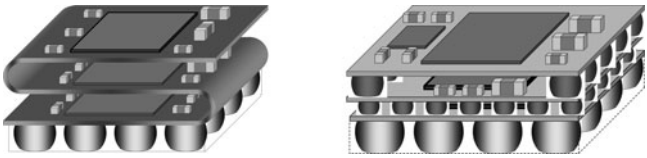


Fig. 1 Two SiPs: integrated on a flexible bended substrate (*left*) and integrated on rigid modules electrically interconnected via conductive solder balls (*right*)

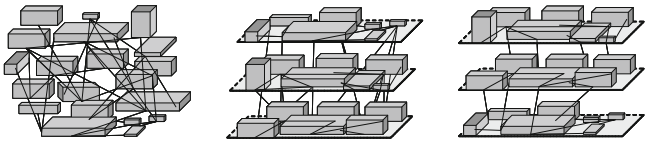


Fig. 2 Abstract view of the layout process of an SiP

Partitioning is to assign the devices to modules. This is currently a manual and time-consuming task. Placement and routing of the devices on the modules can be done with printed circuit board layout tools. However, the tools disregard vertical inter-module connections (VICs). In addition, the engineers select SiP technologies and the number of modules based on experience only.

Hence, we solve partitioning and placement as multiobjective optimization problems and approximate the Pareto-optimal layouts, i.e., layouts where no other layout can be better or at least as good in all objectives. Both height and perimeter of the SiP, the number of VICs, and the wirelength are optimized. For exploring the created layout database we developed a decision support tool 3D SiP Expert. In Sect. 2 we discuss our decision support approach. It streamlines the design with optimization and eliminates time-consuming redesigns by providing a database of alternatives. SiP technologies and the module number are selected by exploring tradeoffs between the objectives.

In Sect. 3 we formulate the placement problem and in Sect. 4 we discuss a constructive heuristic that places the devices on the substrates. It reassembles two intuitive and natural ideas (Chen and Huang 2006): (1) Do less flexible decisions first. (2) Place devices where they fit best. The heuristic does not prove optimality but works effective in practice. We show in Sect. 6 that it is efficient and finds always near-optimal or sometimes even optimal solutions.

We also propose group constraints that implicitly cluster device subgroups, e.g., functional cooperating devices, and that add more structure to the placement. To organize the placement of collaborating devices we need connectivity and topology structure. In Sect. 5 we develop constraints assuring that group devices are connected and satisfy topological properties. We also derive a sufficient condition detecting that a set of group constraints is unsatisfiable.

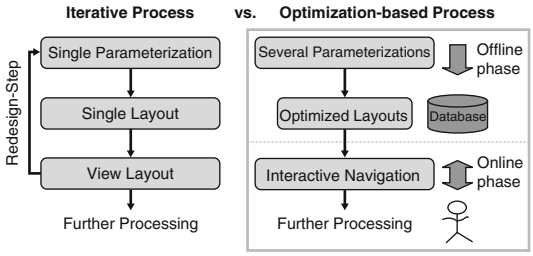


Fig. 3 Current and our novel optimization-based workflow for SiP design

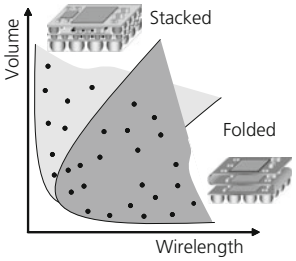


Fig. 4 Comparison of SiP technologies in the objective space

2 Multiobjective Decision Support

The current design workflow is an iterative process. The engineer starts with the electronic schematic and develops a single SiP parameterization, including, e.g., the choice of technology and the module number. A single layout is designed and studied in detail. If it violates a requirement, the engineer must change parameters and has to redesign the layout.

Figure 3 shows the current workflow versus our workflow that concurrently considers alternative parameterizations. The algorithms create a layout database for each parameterization. We subdivide the layout process into two stages: Near Pareto solutions of the partitioning serve as input for the subsequent placement. For each partitioning we create near Pareto placements. The engineer navigates the database interactively and chooses "ideal" layouts by restricting the lower and upper bounds of a selected objective (Richter et al. 2007). These mechanisms guide him to a manageable number of compromise SiP layouts.

Investigating different parameterizations concurrently avoids time-consuming redesign steps. A layout is selected interactively on an objective basis. The engineer directly explores the compromise solutions and compares technologies based on the objective values. Figure 4 shows a two-dimensional (2D) projection of the objective space and the trade-off between technologies whose solutions often cluster in characteristic technology regions. Investigating the technology regions helps the engineer to make an adequate choice.

3 Optimization Problems and Algorithms

In our layout workflow for SiP we have two main optimization problems: partitioning and placement. We neglect the routing because SiP layouts selected with the 3D SiP Expert can be further processed with existing routing tools.

We currently solve the partitioning problem with a Pareto simulated annealing (Jaszkiewicz 2001). In this paper we study the placement problem that is to arrange a set of devices that can be rotated by 90° such that no two devices overlap. Both the module size and the interconnection length between devices of electrical nets should be minimized. Hence, we have a bi-objective problem. However, in this paper we focus on minimizing the module size.

We introduce the following mathematical model: The devices are modeled as rectangles. They are placed in a 2D container that represents a module. We denote $R := \{r_1, \ldots, r_n\}$ as a set of rectangles with index set $I := \{1, \ldots, n\}$; $w_i, h_i \in \mathbb{N}$ represent the width and height, $x_i, y_i \in \mathbb{N}_0$ the coordinates of the lower left corner and $o_i \in \{0, 1\}$ models the orientation of rectangle r_i. $W, H \in \mathbb{N}$ represent the width and height of the container with upper bounds $W_{\max}, H_{\max}$. We use the half perimeter as linear objective for the module size, i.e., $f_1 := W + H$. Therefore, we have the following problem (PP):

$$
\begin{aligned}
\min\ f_1 \qquad & \text{subject to} \\
x_i + s_i^x \le W, \quad & W \le W_{\max}, && (1)\\
y_i + s_i^y \le H, \quad & H \le H_{\max}, && (2)\\
(1-o_i)w_i + o_i h_i = s_i^x, \quad & o_i w_i + (1-o_i)h_i = s_i^y, \ \forall i \in I, && (3)\\
(x_i + s_i^x \le x_j) \vee (x_j + s_j^x \le x_i) & && (4)\\
\vee\, (y_i + s_i^y \le y_j) \vee (y_j + s_j^y \le y_i), & \qquad \forall i, j \in I, i < j.
\end{aligned}
$$

The constraints (1–2) ensure the rectangle containment, (3) define the size of the oriented rectangles and (4) make sure that no two rectangles overlap by arranging them left, right, below or above of each other. In Berger et al. (2008) we developed a constraint program (CP) and a mixed integer program (MIP) for PP. Both programs model the disjunctive non-overlapping constraints by introducing additional variables. In the CP, the variables $C_{ij} \in \{\text{left, right, below, above}\}$ capture the geometric relations between r_i and r_j and propagation algorithms prune the search tree. In the MIP, the variables $z_{ij}^1, z_{ij}^2, z_{ij}^3, z_{ij}^4 \in \{0, 1\}$ are used to model the disjunction with the help of a big-M relaxation (Hooker 2007).

4 Constructive Placement Heuristic

We propose a heuristic that successively chooses a rectangle r_i and its orientation o_i, and places r_i in the container of fixed size (W, H). If the insertion fails, it postpones r_i and retries to place it later. The failed positions are stored to avoid trying

Algorithm 1 Pseudocode of our constructive placement heuristic.

```
while not all rectangles placed and half perimeter ≤ upper bound do
  for all feasible W and H (W ≤ H) that give half perimeter do
    mark all rectangles unplaced.
    while not all rectangles placed and not all positions failed do
      place a rectangle at an insertion position not yet failed.
      if placement fails then
        postpone rectangle and mark insertion position as failed.
      end if
    end while
    if all rectangles placed then
      return placement.
    end if
  end for
  increase half perimeter.
end while
return fail.
```

them again. If r_i cannot be placed at all, the container size is increased and all rectangles are placed again. This procedure is repeated until all rectangles are placed. Algorithm 1 shows the pseudocode of the heuristic. In order to avoid investigating all $(W-1)(H-1)$ possible insertion positions, we only place the rectangles at corner points defined as follows:

Definition 1. A *corner point* p is either an unoccupied corner of the container or a position (x, y) in the container (W, H) where at least two rectangles or one rectangle and the container touch each other in form of a T-junction.

Every rectangle corner can only be involved in at most one corner point yielding at most $4(n+1)$ corner points with the container corners. However, not every placement can be extended by using corner points only. In this case we refine to *intersection points*, i.e., coordinates that align with placed rectangles or any container boundary. There are at most $4(n+1)^2$ intersection points.

We initialize the half perimeter with $f_1 = \left\lfloor 2\sqrt{\sum_{i \in I} w_i h_i} \right\rfloor$ and increment it iteratively. The rectangles, the orientations and the insertion positions are dynamically ordered in every iteration. The basic idea is to prioritize those decisions such that cavities of already placed rectangles are filled as good as possible. This is an intuitive strategy which was already studied in Chen and Huang (2006). We propose the following slightly different evaluation of a placement decision:

Definition 2. Let p be a corner point formed by the placed rectangles r_j and r_k. Let $d^x_{\min}, d^y_{\min}$ be the minimum distance in x and y direction, respectively, of a rectangle r_i placed at p to the placed rectangles of $R \backslash \{r_i, r_j, r_k\}$. The maximal distance of the rectangles R to any container boundary is $d_{\max} = \max(W, H) - \min_{r_i \in R} \min(w_i, h_i)$. Then we define the *cavity factor* as

$$\mathrm{cf} := d_{\max} \min\left(d^x_{\min}, d^y_{\min}\right) + \max\left(d^x_{\min}, d^y_{\min}\right).$$

Lemma 1. *The cavity factor satisfies* $cf \leq d_{\max}\,(d_{\max}+1)$.

Proof. $\min(d^x_{\min}, d^y_{\min}) \leq \max(d^x_{\min}, d^y_{\min}) \leq d_{\max}$ proves the claim.

Hence, we use a *normalized cavity factor*, i.e., $cfn := \frac{cf}{d_{\max}(d_{\max}+1)} \in [0,1]$. In our algorithm we prioritize the rectangle, orientation and insertion position that minimize *cfn*. In case of a tie we choose the larger rectangle, as smaller rectangles can be placed more flexible later on.

Let m be the number of different container sizes which our algorithm tries. Then the runtime complexity of the algorithm is $O(mn^3)$, because for each container size and for each rectangle it has to determine all corner points or all intersection points in $O(n^2)$.

5 Group Constraint Concept

Now we introduce two group constraints for different ways of arranging a group of devices. The examples in Fig. 5 illustrate different basic properties of a placement that can be achieved with the help of the group constraints.

Definition 3. A pair of rectangles (r_i, r_j) is *connected* if they contact each other at their boundaries or corners. An undirected graph G is called the *connection graph* of rectangles $R' \subseteq R$ if its vertices represent R' and there is an edge between vertices r_i and r_j if the rectangles are connected. A group of rectangles R' is *connected* if G is connected. A *connected group constraint* $C_C\,(R')$ requires that the rectangles R' are a connected group.

Definition 4. An *orthoregion* D is the union of finitely many rectangles $R' \subseteq R$, i.e., $D = \bigcup_{r \in R'} r$. The *box hull* $bh(D)$ is the smallest rectangle that contains D. A connected group of rectangles R' is *box-free* if no rectangle of $R \backslash R'$ intersects with $bh(D)$. A *box-free group constraint* $C_B\,(R')$ is a connected group constraint requiring the rectangles R' to be a box-free group.

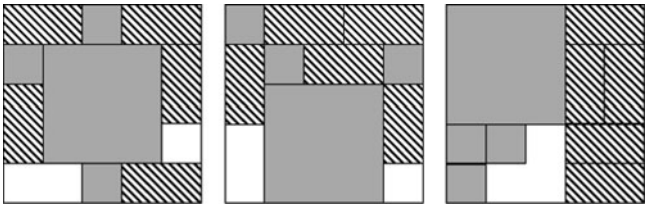


Fig. 5 Placements with five hatched devices of a collaborating group: random arrangement (l); contacting each other (m); exclusively arranged within a rectangle (r). The four *gray* devices are part of no group and the *white* regions are unoccupied

In the left placement of Fig. 5 the hatched rectangles are disconnected but in the middle they are a connected group. Only in the right placement the hatched rectangles form a box-free connected group.

Lemma 2. *The box-free group constraint network with the constraints* $C_B^1 := \{r_1, r_2\}, C_B^2 := \{r_2, r_3\}$ *and* $C_B^3 := \{r_3, r_1\}$ *is unsatisfiable.*

Proof. We assume C_B^1, C_B^2, C_B^3 are satisfied. Hence, r_1, r_2, r_3 must be connected and contact each other in exactly one point (x, y). However, then one box hull of two rectangles is not free of the third rectangle in (x, y).

Now we sketch the approach to integrate connected and box-free group constraints in our heuristic and in the CP and MIP model of Berger et al. (2008). For a connected group constraint we store its connection graph and test if the graph is connected. A box-free group constraint is transformed into constraints on the coordinates and the geometric relations of the rectangles.

Lemma 3. *If all rectangles* R' *of a box-free group constraint* $C_B(R')$ *are connected and in the same geometric relations to* $R \setminus R'$ *then* $C_B(R')$ *is satisfied.*

Proof. We assume that $C_B(R')$ is violated and that there exists no rectangle of $R \setminus R'$ that is in different geometric relations to two rectangles of R'. Then we can find two horizontal and two vertical lines that separate R' from those rectangles of $R \setminus R'$ that are left, right, below and above of R'. Consequently, the box hull of R' is free of other rectangles which contradicts our assumption.

However, when at least two box-free group constraints are non-disjoint we have to use constraints on the coordinates instead of the geometric relations.

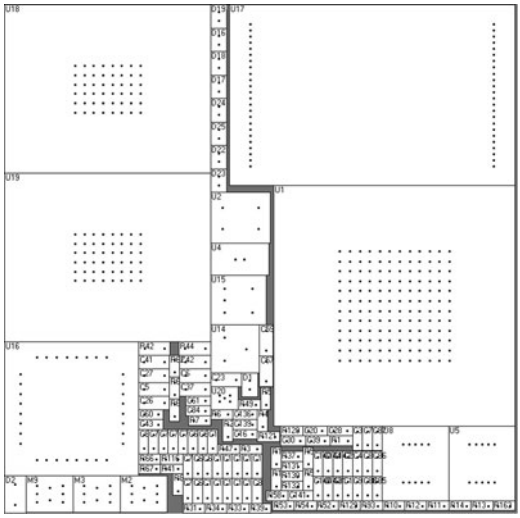


Fig. 6 Real world SiP placed by our heuristic

6 Computational Results

We implemented the constructive heuristic with ILOG Solver 6.6 and tested instances of size $n = 5, \ldots, 27$ where the rectangle sizes are inspired by SiP devices. The instances can be found in Berger et al. (2008). Also, we run the algorithm on a real world SiP with $n = 133$ devices and place them on a single module.

The second column of Table 1 shows that our algorithm runs within seconds. The third column shows the gap to the optimal solutions we published in Berger et al. (2008) (dash if optimality is unknown). All solutions are within 6% of optimality. The fourth column lists the percentage of lost module area. A qualitative placement has less than 15% of free area in practice. Therefore, our solutions are of high quality. The SiP shown in Fig. 6 was found in 11 CPU seconds.

Figure 7 illustrates placements where no (left), a connected (middle) and a box-free group (right) constraint for the dark gray group of devices is applied. It shows that the constraints introduce more structure for collaborating device groups. More structured placements are preferred by the engineers.

Table 1 Results for the heuristic

n	CPU (sec.)	Gap (%)	Lost area (%)
5	0.05	0.00	1.79
6	0.02	0.00	7.83
7	0.03	2.25	11.34
8	0.03	1.12	7.41
9	0.02	1.14	6.77
10	0.08	5.45	10.64
11	0.09	–	11.82
12	0.05	–	5.98
13	0.14	1.59	7.89
14	0.06	–	4.59
15	0.28	2.37	5.14
16	0.20	–	7.65
17	0.23	–	4.40
18	0.20	–	6.21
19	0.25	–	6.49
20	0.39	–	8.42
21	0.36	–	7.74
22	0.30	–	4.48
23	0.39	–	6.57
24	0.39	–	5.92
25	0.44	–	5.76
26	1.09	–	8.58
27	0.64	–	6.26

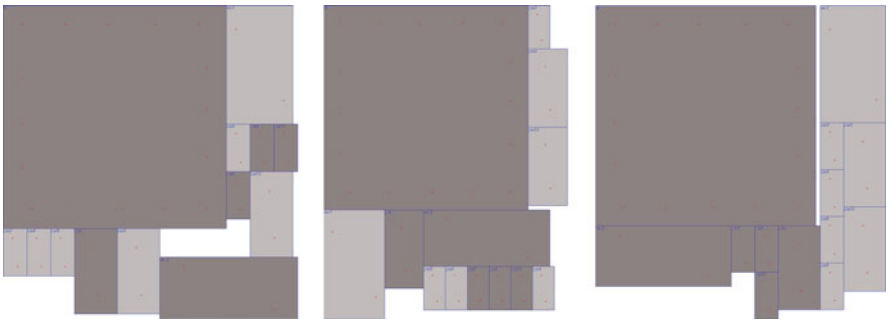


Fig. 7 Group constraints applied to the test instance with $n = 13$

7 Conclusion

We presented an optimization approach for improving both the basis for decision-making and the automation in 2.5D SiP design. This is confirmed by engineers that tested our tool. We developed a heuristic that produces high quality placements fast. Finally we introduced group constraints that help to place collaborating device groups. In future work we will extend the heuristic for bi-objective placement and refine the group constraint concept.

Acknowledgements We are indebted to two anonymous referees for their valuable comments. This work is funded by the Fraunhofer ITWM.

References

Berger, M., Schröder, M., & Küfer, K.-H. (2008). *A constraint programming approach for the two-dimensional rectangular packing problem with orthogonal orientations.* Technical Report 147, Fraunhofer ITWM.

Chen, D., & Huang, W. (2006). A novel quasi-human heuristic algorithm for two-dimensional rectangle packing problem. *International Journal of Computer Science and Network Security, 6*(12), 115–120.

Hooker, J. N. (2007). *Integrated Methods for Optimization*. Berlin, Heidelberg, New York: Springer.

Jaszkiewicz, A. (2001). *Multiple objective metaheuristic algorithms for combinatorial optimization*. Habilitation Thesis, Poznan University of Technology.

Polityko, D. D. (2008). *Physikalischer Entwurf für die Vertikale SiP Integration*. PhD Thesis, Berlin Institute of Technology.

Richter, C., Polityko, D. D., Hefer, J., Guttowski, S., Reichl, H., Berger, M., Nowak, U., & Schröder, M. (2007). Technology aware modeling of 2.5D-SiP for automation in physical design. In *Proceedings of the 9th Electronics Packaging Technology Conference* (pp. 623–630).

Multi-Objective Quality Assessment for EA Parameter Tuning

Heike Trautmann, Boris Naujoks, and Mike Preuss

Abstract Evolutionary algorithms are non-deterministic and highly parameterizable optimization methods. Therefore, the setting of parameters greatly influences their performance and methods for parameter tuning became more and more popular in recent years. However, obtained parameter settings are usually valid only for the tackled combination of algorithm, problem, and performance measure. In most investigations concerning EA tuning, only one performance measure is utilized, inherently defining it as 'user preference'. However, users' preferences may be different. While one user may only look for a single best solution from a couple of runs for a design problem (single-excellent case), another may be interested in a generally stable behavior of the algorithm displayed by a good expected value of multiple runs and a low variance (robust case). An efficient handling of this trade-off is investigated here. In particular, we investigate the possibility to control the behavior of a given algorithm on a given problem between the two stated extremes via changing one or a small number of parameters.

1 Introduction

Industrial optimization is generally guided by searching for the global best solution. However, today's technology enables engineers to model highly complex processes and thus make them accessible for optimization based on computer experiments. The resulting optimization problems are highly complex as well, and solving such tasks, i.e. to find the global best solution, is hard.

To address such problems, randomized search heuristics such as evolutionary algorithms (EA, cf. Eiben and Smith 2003; DeJong 2006) have been developed. They are able to explore the search space much better than deterministic search strategies such as e.g. gradient based methods.

B. Naujoks (✉)
Log!n GmbH, Schwelm, Germany
e-mail: Boris.Naujoks@login-online.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_88,

Due to the stochastic nature of such algorithms, different optimization runs yield different qualities of the results. While it is advantageous to have good average results to measure the quality of some optimization procedure, a single very good solution is more appealing to the engineer who solely looks for the best solution to implement.

In this paper we are addressing this trade-off by tuning EA accordingly. To this end, the framework of Sequential Parameter Optimization (SPO) (cf. Sect. 3.1 or Bartz-Beielstein 2006 and Jones et al. 1998) is used to optimize EA parameter settings for good average solutions on the one hand and single excellent solutions on the other hand. This ends up to be a multi-objective optimization problem, and we search for parameters to run along the resulting Pareto front enabling the user to scale from one optimization goal (e.g. good average solutions) to the other.

The paper is organized as follows: the next section gives a small introduction to test function sets for Evolutionary Multi-Objective Optimization algorithms (EMOA, cf. Deb 2001; Coello Coello et al. 2002) and motivates as well as introduces the new combination of single-objective test functions into a multi-objective one. Section 3 describes the experiments executed and analyses the results. Section 4 concludes the contribution and provides an outlook on further research.

2 Definition of Test Functions

Since research on EMOA established a rather popular scientific field in the late 1990s, different sets of test functions to compare the quality of different approaches have been proposed. The first frequently used such set was proposed by Zitzler et al. (2000). Some preliminary studies for the contribution at hand have been carried out on function ZDT4. Although these test functions are still used frequently, they feature several disadvantages, the most important one is sharing only two objectives. This was overcome by publishing the DLL test function set (cf. Deb et al. 2002).

Up to now, the test functions from above are still in use, however they have been improved and adjusted to fit more general situations. The most current and frequently used function set was published by Huband et al. (2006). Some of these test functions have been invoked to compare different EMU approaches during one of the latest conferences of the *evolutionary computation* (ESC) community. The congress on evolutionary computation (CEO) 2007 test function set (Huang et al. 2007) is the state-of-the-art for EMU algorithm comparisons today.

Preliminary results on different test functions revealed that typical EMU test functions such as ZDT4 from Zitzler et al. (2000) or SYMPATHY from the CEO 2007 collection are too easy to solve for the considered algorithm, e.g. SUMS-EMOA (cf. Sect. 3.1). Consequently, more complex test functions from the field of single-objective optimization have been considered. Surprisingly, this work is the first approach to combine such single-objective test functions to design multi-objective ones to our knowledge.

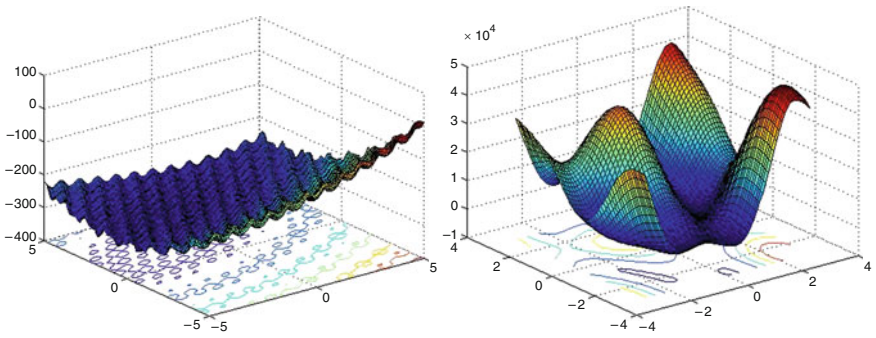


Fig. 1 Astringing's (F10, *left*) and Schlemiel's test function (F12, *right*) for a two dimensional search space

The two single objective test functions considered were incorporated in a comparison on a former CEO conference, i.e. CEO 2005 (Suganthan et al. 2005). We decided to combine the shifted and rotated version of Astringing's function (F10) from the above collection with Schlemiel's function (F12) to receive F = (F10, F12). Not to start with too complex situations, it was decided to start with a small search space dimension of $d = 2$. The next step will be a generalization to higher dimensional cases. A plot of both functions for this search space is provided in Fig. 1.

3 Experiments and Results

3.1 Experiments

As mentioned above, the evolutionary multi-objective optimization algorithm used for this paper is the SUMS-EMOA (Beume et al. 2007). This algorithm considers two ranking criteria to select the succeeding population within a $(\mu + 1)$-selection scheme[1], namely the dominance rank and selection based on the hypervolemia. The hypervolemia is the space covered by the Pareto front with respect to some predefined reference point (cf. Zitzler 1999). While the dominance rank is a standard selection technique as used in popular algorithms like NASA-II (cf. Deb et al. 2002) as well, the SUMS-EMOA in particular rejects the individual contributing the least hypervolemia to the worst ranked Pareto front.

The standard variation operators for the generation of new individuals in the EMOA field are simulated binary crossover (SAX, cf. Deb 2001) and polynomial

[1] cf. standard EA literature for the notation. The general $(\mu + \lambda)$-selection scheme indicates that μ parents exist within the EA and λ offsprings are generated in each generation. The succeeding population is formed by the μ best solutions from the union of parents and offsprings.

Table 1 SUMS-EMOA settings for SPO

	μ	P_C	η_C	P_M	η_M
Interval	[10, 200]	[0.1,1]	[5, 50]	[0.1,1]	[5, 50]
Standard	100	1	20	1/n	15

The indices C and M refer to the crossover and the mutation operator, μ equals the population size

mutation (ibid.). Both are typically controlled by two parameters, namely their application probability and a parameter controlling the deviation of the underlying distribution. All parameters considered within our investigation are listed together with the default values (standard) and the interval we employed for the investigation (interval) in Table 1.

The framework of Sequential Parameter Optimization (SPO, Bartz-Beielstein 2006) is a stepwise procedure to indicate good parameter settings for one algorithm applied to a special application or test function. This stepwise procedure starts with a space-filling design [e.g. Latin Hypercube Design (LHD)] within the algorithm parameter space and evaluates the algorithm's performance for the corresponding parameter settings. Building sophisticated models based on the outcome of these first experiments, regions of interest are determined within the parameter space and the search for good parameter settings is continued focusing on these regions.

For each optimization run of SUMS-EMOA 2,500 evaluations have been executed with 2,000 runs allowed for each SPO. The number of design point evaluations was set to 10 for the first experiments.

3.2 Results

First experiments have been conducted on the function SYMPATHY from the CEO 2007 test case collection, but this problem turned out to be too easy. As a result, no significant trade-off with respect to the hypervolemia could be observed.

Further preliminary experiments have been conducted on ZDT4 from the corresponding test case collection. Here, a possible influence of two parameters, namely the population size (μ) and the recombination probability (P_C) was observed. However, the drawback was that not enough points on the Pareto front were generated during the SPO runs. Furthermore, the estimation of the Pareto front was not reliable enough.

3.2.1 Results on F10–F12

As a more complex test case, the combination of the two single-objective test functions F10 and F12 was considered. Comparing all different parameter influences with each other in a scatter plot, a possible influence of the recombination

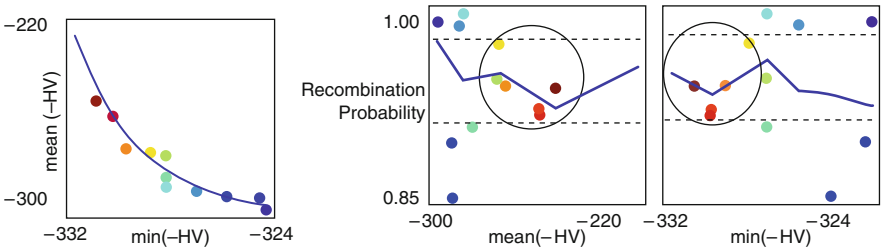


Fig. 2 1st and 2nd set of non-dominated points. The original goal was to find a scalable input parameter to move along the estimated Pareto front (*left*). If any of the parameters invoked and compared in a scatter plot, the recombination probability could have a possible scaling effect (*right*)

probability was identified. The corresponding diagrams have been merged in Fig. 2, right part, where the comparison of the recombination probability and both hypervolemia measurements (mean(HV) and max(HV)) is presented. Here, the scaling effect can be observed.

The left part of Fig. 2 shows the trade-off between the average hypervolemia values and the best hypervolemia values in general. Here and for all following figures, the Pareto-front was incorporated next to the solutions dominated by only one solution (rank two Pareto-front) resulting in a 'Pareto front region'. Thus a kind of ε-dominance approach is carried out being aware of an uncertainty in the location of each point of the front.

As a consequence of the identification of P_C, a grid test was conducted for this parameter. Here, all other parameters under investigation were fixed at their median level. The results achieved from this grid test are given in Fig. 3. The colors indicate different areas of the Pareto front. Unfortunately the scaling only focused on the upper part of the Pareto front which indicated that important parameter interactions have been neglected (cf. Fig. 3, left part).

As the visual analysis of the scatterplot-matrices thus was not promising, statistical classification methods like CART (Classification and Regression Tree) seemed to be more appropriate to at least discriminate between the two "edges" of the Pareto front (Fig. 3, right part). For this purpose the reliability of the approximation of the "true" Pareto front is a key issue as the desired analysis only makes sense if the uncertainty of the estimated points is sufficiently low.

Figure 4 shows all 20 hypervolemia values for each point of the Pareto front and reveals the high variance of the SUMS-EMOA performance for all parameter settings. Certainly this is due to the stochastic nature of an evolutionary algorithm. Nevertheless, the extent of the variation is surprisingly high. In order to reduce the variance, the arithmetic mean was replaced by the median as a robust estimator of average performance. Though this seemed to be a promising approach as the arithmetic mean is highly influenced by the maximum hypervolemia value, the spread of solutions does not differ significantly from the situation before.

A validation of the points is presented in Fig. 5, left part. The SUMS-EMOA was run 100 resp. 1,000 times using the input parameter combination of each point to be

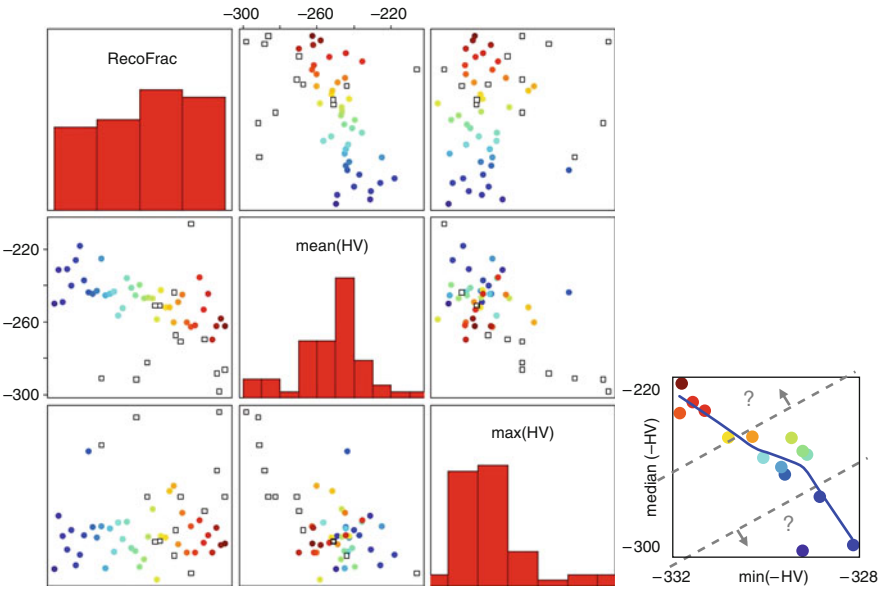


Fig. 3 Grid test for influence of recombination probability (*left part*) and 1st and 2nd set of non-dominated points for F10–F12. The research goal changed to at least perform a classification regarding the "edges" of the Pareto front (*right part*)

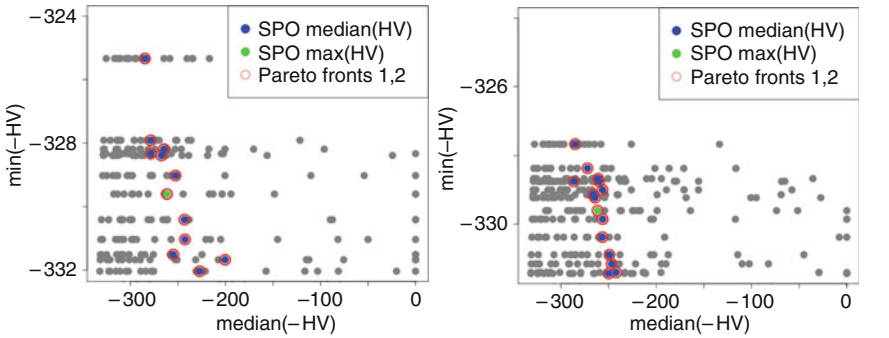


Fig. 4 1st and 2nd non-dominated points supplemented by the 20 single repeats for each configuration (*grey*). *Left*: SPO Mean(HV), *Right*: SPO Median(HV)

investigated, and the minimum as well as the median of the resulting hypervolemia indicator was computed. Each validation run is marked by a different symbol, and the colors match the respective point on the original estimated front. Convex hulls of the points are used to roughly visualize the region of the point locations. It becomes obvious that the Pareto front estimation so far is not stable at all, and that the underestimation of the solution quality increases with decreasing number of SUMS-EMOA runs.

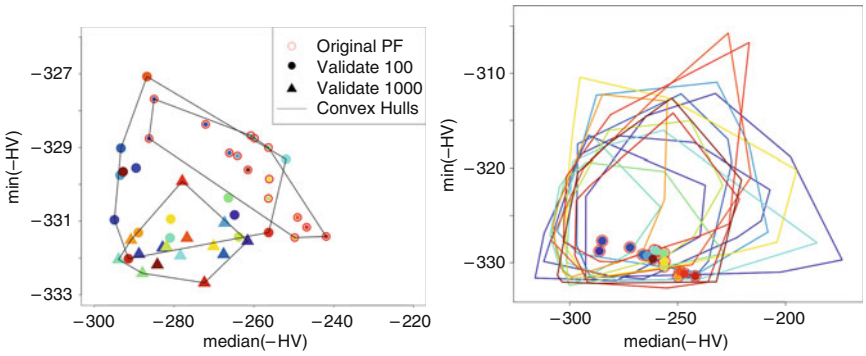


Fig. 5 Validation of points on 1st and 2nd non-dominated front by repeated runs of SUMS-EMOA for each underlying parameter setting (*left*). About 50 runs of 20 validations of the points are performed and their convex hulls are plotted in matching color (*right*)

In order to measure the uncertainty of the Pareto points more accurately, in a second validation step the 1,000 repeated SUMS-EMOA runs were treated as 50 separate validations with in each case 20 SUMS-EMOA runs. For each of the 50 validations the two objectives $max(HV)$ and $median(HV)$ were computed resulting in 50 validation points for each Pareto point. Convex Hulls of these points visualize the spread of the validation points in the objective space (Fig. 5, right part). The colors of the lines match the color of the respective Pareto front point. Once again the high uncertainty of the Pareto front estimation becomes obvious while the variance in the median values exceeds the variance in the single excellent solutions.

4 Conclusions and Outlook

The original task was to find one or two algorithm parameters as handles for switching between a good average and a good peak performance. We started by tuning parameters of the investigated SUMS-EMOA regarding these two performance goals, hoping to find a Pareto front between them. It turned out that the algorithm configurations representing the front behave much more stochastic than expected. Consequently, all trials of identifying parameters as switches failed, and the obtained Pareto fronts are proven to be rather questionable. These result led to very interesting insight and more 'informed' investigations that shall be carried out in future. It can be stated that our original research goal was too ambitious due to the high uncertainty of the desired Pareto front estimation. However, the awareness of this 'failure' itself is an important research result leading to a better understanding of EMOA results.

In order to try to reduce the solution variance, next steps will include an increase in the maximum number of function evaluations of the SUMS-EMOA, which will result in a higher convergence probability. Secondly, it will be investigated how the

two separate univariate SPO runs can efficiently be replaced by a multivariate SPO technique Ponweiser et al. 2008 in combination with specific techniques for noisy environments.

References

Bartz-Beielstein, T. (2006). *Experimental research in evolutionary computation – The new experimentalism*. Natural Computing Series. Berlin: Springer.

Beume, N., Naujoks, B., & Emmerich, M. (2007). SMS-EMOA: Multiobjective selection based on dominated hypervolume. *European Journal of Operational Research, 181*(3), 1653–1669.

Coello Coello, C. A., Van Veldhuizen, D. A., & Lamont, G. B. (2002). *Evolutionary algorithms for solving multi-objective problems*. New York: Kluwer Academic Publishers. ISBN 0-3064-6762-3.

Deb, K. (2001). *Multi-objective optimization using evolutionary algorithms*. Chichester, UK: Wiley.

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA–II. *IEEE Transactions on Evolutionary Computation, 6*(2), 182–197.

Deb, K., Thiele, L., Laumanns, M., & Zitzler, E. (2002). Scalable multi-objective optimization test problems. In *Congress on Evolutionary Computation (CEC 2002)* (pp. 825–830). Piscataway, New Jersey: IEEE Press.

DeJong, K. A. (2006). *Evolutionary computation: A unified approach*. Cambridge, MA: MIT Press.

Eiben, A. E. & Smith, J. E. (2003). *Introduction to evolutionary computing*. Natural Computing Series. Berlin: Springer. ISBN: 3-540-40184-9.

Huang, V. L., Qin, A. K., Deb, K., Zitzler, E., Suganthan, P. N., Liang, J. J., Preuss, M., & Huband, S. (2007). *Problem definitions for performance assessment of multi-objective optimization algorithms.* Technical report, Nanyang Technological University, Singapore.

Huband, S., Hingston, P., Barone, L., & While, L. (2006). A review of multiobjective test problems and a scalable test problem toolkit. *IEEE Transactions on Evolutionary Computation, 10*(5), 477–506.

Jones, D., Schonlau, M., & Welch, W. (1998). Efficient global optimization of expensive black-box-functions. *Journal of Global Optimization, 13*, 455–492.

Ponweiser, W., Wagner, T., Biermann, D., & Vincze, M. (2008). Multiobjective optimization on a limited budget of evaluations using model-assisted s-metric selection. In G. Rudolph et al. (Eds.), *Parallel problem solving from nature (PPSN)* (pp. 784–794). Berlin: Springer.

Suganthan, P. N., Hansen, N., Liang, J. J., Deb, K., Chen, Y.-P., Auger, A., & Tiwari, S. (2005). *Problem definitions and evaluation criteria for the cec 2005 special session on real-parameter optimization.* Technical report, Nanyang Technological University, Singapore, and KanGAL Report 2005005, IIT Kanpur, India.

Zitzler, E. (1999). *Evolutionary algorithms for multiobjective optimization: Methods and applications*. PhD thesis, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland.

Zitzler, E., Deb, K., and & Thiele, L. (2000). Comparison of multiobjective evolutionary algorithms: Empirical results. *Evolutionary Computation, 8*(2), 173–195.

A Novel Multi-Objective Target Value Optimization Approach

S. Wenzel, S. Straatmann, L. Kwiatkowski, P. Schmelzer, and J. Kunert

Abstract In the recent years the Efficient Global Optimization algorithm (EGO) by Jones has become a widely used technique in engineering applications. Based on a small initial design a surrogate model is fitted and updated sequentially using the so-called expected improvement criterion to find the global optimum. Henkenjohann and Kunert presented a multivariate extension of EGO using the concept of desirabilities. Due to difficulties with the distribution of the desirability index, their extension is restricted to one-sided desirabilities only. We therefore extended their strategy to enable target value optimization using two-sided desirabilities. Instead of calculating the exact expected improvement using the uncertainty distribution we determine the improvement based on a very rough Monte Carlo simulation. A case study from mechanical engineering demonstrates the usability of the approach.

1 Introduction

The optimization of an engineering process usually is very complex. Large numbers of influencing parameters and several objectives are to be optimized simultaneously. At the same time, engineers try to reduce the number of experimental runs needed due to time and cost constraints. Sequential optimization has become a very popular tool for such situations. The most popular algorithm is the Efficient Global Optimization algorithm (EGO) by Jones et al. (1998). Since then various variants of EGO for multivariate problems or constraint problems appeared. All those algorithms are designed for minimization or maximization problems. We were asked to optimize a simple pot produced with the sheet metal spinning process. This is an incremental forming process that produces complex rotationally symmetric workpieces (cf. Fig. 1).

The geometry of the mandrel defines the final shape of the workpiece. We decided to maximize the sheet thickness and to minimize the other objectives.

S. Wenzel (✉)
Department of Statistics, Technische Universität Dortmund, Dortmund, Germany
e-mail: wenzel@statistik.tu-dortmund.de

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_89,

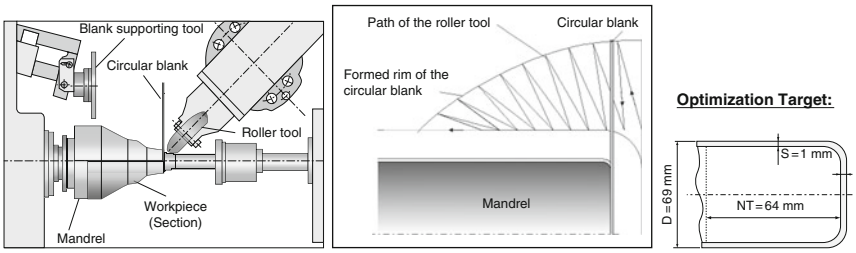


Fig. 1 Producing a pot with the sheet metal spinning process

Unfortunately, unexpected swelling occurred and the optimal pot missed the actual targets. In this particular application an optimization regarding a specified target value is needed. The existing algorithms can not be transformed straight forward into a target-value problem, since a simple transformation of the objective into an objective that can be minimized or maximized yields in difficult and mostly unknown distributions of the objective. This paper introduces a multivariate sequential optimization algorithm that is able to handle target value problems. It is structured as follows. Section 2 gives a brief introduction to the EGO algorithm, the expected improvement criterion, and a multivariate adaption of EGO using desirabilities. In Sect. 3, the new approach to target-value optimization is presented. Section 4 discusses the presence of unknown constraints. Optimization results for the spinning process are presented in Sect. 5. Finally, in Sect. 6 a short conclusion is given.

2 Efficient Global Optimization (EGO)

The EGO algorithm was introduced by Jones et al. (1998). EGO starts with an initial design to fit a surrogate model. With the help of the expected improvement criterion (EI) so-called updating points are chosen to refine the surrogate model until a defined stopping criterion is reached. The EI balances the exploitation from the surrogate model, where the prediction is optimal with the need for exploration where the uncertainty is high. The EI of x given the observations $y^{(n)}$ and the current optimum observation y_{max} is the expected value of the improvement $I(x) = 1_{[Y(x)>y_{max}]}(Y(x) - y_{max})$ conditional on the uncertainty distribution of the model $f_{Y(x)}$, hence

$$E[I(x)|y^{(n)}] = \int_{y_{max}}^{\infty} (I(z))\, f_{Y(x)|y^{(n)}}(z)dz. \tag{1}$$

The improvement $I(x)$ thereby addresses the potential in optimization for x. The point x that maximizes the EI is the new updating point. The refinement process is stopped as soon as the maximum EI is smaller than 1% of y_{max}. Jones et al.

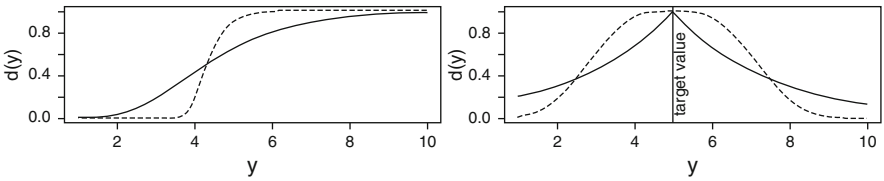


Fig. 2 Exemplary one- (*left*) and two-sided (*right*) Harrington desirability functions

suggest to use Kriging models as surrogate models. For those models the uncertainty distribution is $Y(x)|y^{(n)} \sim N(\hat{y}(x), s^2_{\hat{y}(x)})$ and the EI can be given in a closed form Jones et al. (1998).

The concept of desirability

Harrington (1965) introduced the concept of desirability to handle multivariate problems. Desirability functions are used to map the objectives to the interval $[0, 1]$. The transformation enables the comparison of objectives with different scales. Figure 2 shows different examples to specify a desirability function. On the left hand side one-sided specifications for a maximization problem are given. The desirability starts at zero indicating the results that are unacceptable. The desirability $d(y)$ increases with increasing result value y and stays 1 when a desired value is exceeded. In the same way a minimization problem can be specified. The right hand side of the figure shows two-sided specifications that account for target value problems. The desirability is 1 when the target value is met and gets smaller the further away the results move from the target. Different kurtosis values can be used to influence how strong divergences are penalized. Derringer and Suich (1980) introduced asymmetric specifications, that allow to penalize differently whether the target is exceeded or under-run. Having transformed each objective with a desirability function, they are joined to a univariate desirability index. It usually is the (weighted) geometric mean of the desirabilities. The maximum of the index indicates where the joint optimum is situated. Desirability indices using two-sided desirability functions exactly addresses the issue of target-value optimization.

Multivariate expected improvement using desirabilities

Henkenjohann and Kunert (2007) introduced a multivariate expected improvement criterion using the concept of desirabilities. This EI is based on the vector of desirabilities $d(x)$ and the index $DI(x)$:

$$E[I(x)|y_1^{(n)}, \ldots, y_m^{(n)}] = \int\limits_{\substack{z \in [0,1]^m: \\ DI(z) > DI_{max}}} I(z) \cdot f_{d(x)|y_1^{(n)},\ldots,y_m^{(n)}}(z) dz, \qquad (2)$$

where $I(x) = 1_{[DI(x) > DI_{max}]} \min\limits_{f \in F_{max}} (||f - DI(x)||)$ and $f_{d(x)|y_1^{(n)},\ldots,y_m^{(n)}}$ is the joint distribution of the vector of desirability functions. The joint distribution of

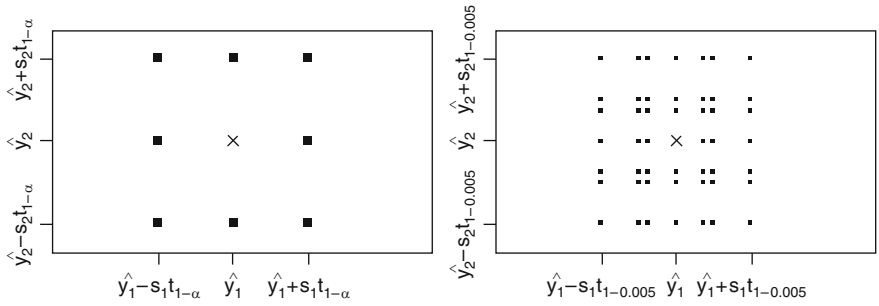


Fig. 3 Virtual observations for $\alpha = 0.005$ (*left*) and $\alpha = (0.1, 0.05, 0.005)$ (*right*)

the desirability vector is very hard to determine and could so far only be derived for the special case of one-sided Harrington desirability functions. For length of computation time, a full Monte-Carlo Simulation of the distribution is also not feasible. Hence, this approach cannot handle target-value problems.

3 The New Approach

On the basis of the multivariate expected improvement of Henkenjohann and Kunert (2007) we present an optimization heuristic to target-value that uses two-sided desirabilities. Instead of deriving or fully simulating the distribution of the desirability vector we calculate only a very rough approximation of the uncertainty distribution of the untransformed predictions by so-called virtual observations. Similarly to Cox et al. (1997) we construct virtual observations by means of $1 - \alpha$ confidence boundaries for the predictions $\widehat{y_i}$ using the prediction errors s_i. Figure 3 shows for one exemplary predicted point $(\hat{y}_1(x), \hat{y}_2(x))$ the upper and lower confidence boundary for each objective y_1 and y_2. This results in nine virtual observations, including the prediction, each representing one possible true result $(y_1(x), y_2(x))$. The virtual observations give a very rough impression of how the model uncertainty influences the prediction. If several confidence levels are used at the same time, e.g. three levels as in the right hand side of the figure, the impression gets better. A large number of levels of course would yield in a full Monte-Carlo simulation of the model uncertainty distribution. But already for three levels 48 virtual observations have to be constructed for each predicted point, and the following procedure would become inefficient. Figure 4a shows virtual observations of two confidence levels for a whole vector of predictions. For better illustration, the example has two objectives and only one influencing parameter. The predictions are quite certain in the interval [1, 5] and very uncertain in the interval [5,11]. The target is drawn with a horizontal line. The current local optimum for y_1 lays in the space with small model uncertainty. Whereas the predictions do not indicate any improvements in EI in the uncertain area, the virtual observations show a high potential of containing global

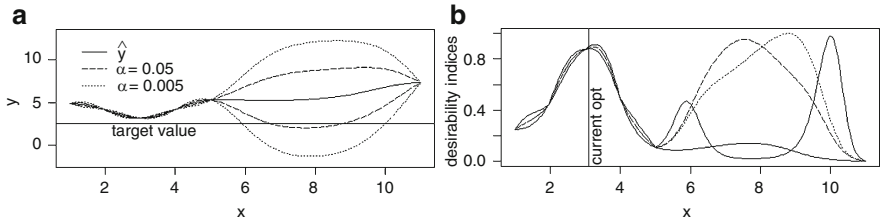


Fig. 4 Exemplary virtual observations for two α-levels and resulting desirability indices

optima in different places. The same applies to all other objectives. To determine the possible joint global optima of the multivariate problem, we now use desirabilities to transform the problem into a univariate one. The vectors of virtual observations coming from the different α levels are transformed to (two-sided) desirability vectors. The desirability vectors of all objectives are then cross-combined with each other, resulting in (number of α-levels)$^{\text{number of objectives}}$ different indices. Figure 4b shows some indices for the simple example, we skip those having the maximum in the already found local optimum. Some of the indices have new peaks that are larger than the peak of the current local optimum. Those are points with a high potential for a joint global optimum. For each desirability index improvements are calculated, resulting in a vector of improvement values per desirability index. We determine the points that maximize one of the vectors of improvement values and get a set of points that we call the candidates for global optimums. Since some of the candidates often lie close to each other or occur several times, we use standard hierarchical clustering to determine groups of points. The model is refined with the candidates which are the most in the middle of those groups.

3.1 Exemplary Progress

Figure 5 shows an exemplary progress of the set of candidates. The example has three objectives and two influencing factors. The contour plot of the true desirability index shows two global and one local optima. In the beginning of the optimization when the model is still very uncertain the candidates are scattered widely. During the further steps, the candidates concentrate more and more on the global optima, but also the region of the local optimum is tested. In the end, the two global optima of the true desirability index are found.

3.2 Stopping Criterion

In contrast to Jones et al. we have several improvement vectors, and our stopping criterion has to take them all into account. With extensive studies and comparisons

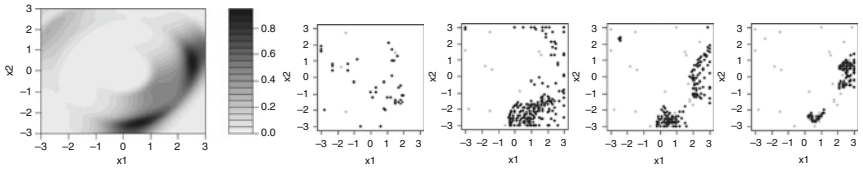


Fig. 5 Exemplary progress of the set of candidates during the optimization steps

we found that the best stopping indicator is taking the maximum of all vectors of improvements and choose again the maximum out of it. The algorithm can be stopped, when the maximum of the maximums reaches 0, i.e. no virtual observation reaches better results than the current found. Depending on the α levels it will take a very long time until 0 is reached, therefore we propose to observe the progress of the criterion and stop when it saturates at a level close to 0.

4 Handling of Missing Values

The optimization of engineering processes often involves the problem of unknown failure boundaries. If an updating point produces a failure part, the quality measures are missing. Ignoring them would be fatal, since the model stays unchanged in this region and the algorithm will stall there. Forrester et al. (2006) therefore suggest to substitute the failure point with a penalized prediction that diverts the algorithm towards the feasible region. For the two-sided case, we suggest to substitute the failure at x with $\hat{y}(x) - s(x)$ if the prediction of the last surrogate model $\hat{y}(x)$ is smaller than the target value and $\hat{y}(x) + s(x)$ if $\hat{y}(x)$ exceeds the target value.

Additionally, we adapt the strategy of Henkenjohann et al. (2005) to exclude failure regions from the parameter space. Assuming the feasible area of the process is convex, the area lying from the viewpoint of a feasible point behind a failure point must belong to the failure region (cf. left hand side of Fig. 6). To determine failure regions a polyhedral convex cone is spanned by the feasible points and a failure point as cone tip with the very efficient double description method by Fukuda and Alain (1996). All points lying inside this cone belong to the failure region and can be excluded from the parameter space. The right hand side of Fig. 6 shows a parameter space and the excluded regions as white space. The exclusion of known failure regions prohibits that the algorithm suggests points in an uncertain area that lies in the failure region.

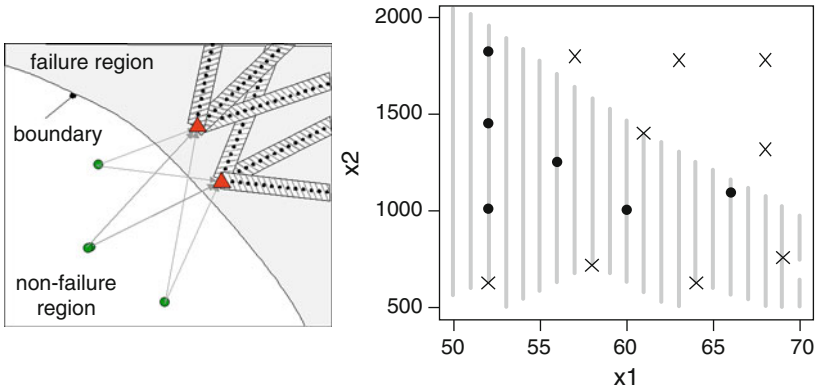


Fig. 6 Restriction of the parameter space with polyhedral convex cones (Henkenjohann and Kunert 2007)

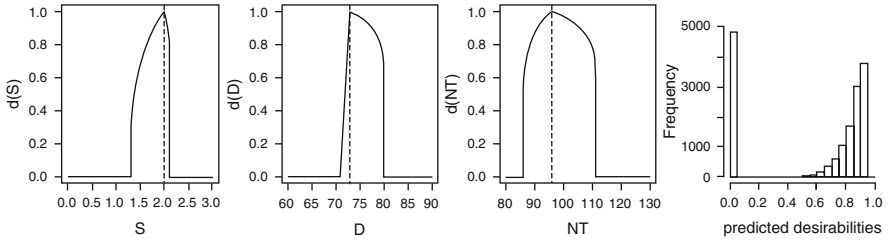


Fig. 7 Desirability functions for the sheet metal spinning process

5 Case Study

We tested our algorithm with the optimization of the sheet metal spinning process (cf. Sect. 1). Again a pot is optimized over the three objectives, depth NT, diameter D, thickness S all measured in mm. The optimization deals with six influencing parameters (A–F) and a grid of 15,625 points representing settings in the six-dimensional parameter space. Figure 7 shows the specified asymmetric two-sided Derringer-Suich desirabilities. Note that for the thickness, results that exceed the original sheet thickness of 2 mm, are penalized very strongly to prohibit swelling of the material. The algorithm started with a space-filling design with only 15 points and used three confidence boundary levels 0.0001, 0.001 and 0.01. The global optimum is found within six steps, where the computing time per step was about 30 min on a 2 Ghz machine. The last part meets the target values ($S = 2$ mm, $D = 73$ mm and $NT = 96$ mm) very good, especially the thickness S does not exceed the target as before.

Table 1 shows for each step the currently found optimum and its corresponding desirability index. The desirability could be improved from about 0.93 to 0.97. Although, this seems not to be a great improvement, the algorithm did a good job.

Table 1 Current optima in the single steps of the optimization

Points	DI	A	B	C	D	E	F	*S*	*D*	*NT*
1–15	0.9295449	−1	10	30	0.95	0	1.5	1.74	76.4	97.88
1–25	0.9417639	−1	20	38	1.01	0	1.5	1.86	76.9	94.1
1–32	0.9417639	−1	20	38	1.01	0	1.5	1.86	76.9	94.1
1–39	0.9436689	1	20	54	0.92	−1	2.7	1.96	76.5	91.6
1–46	0.9513387	2	20	62	0.92	−2	1.5	1.86	75.9	94.23
1–51	0.9719737	2	10	30	0.92	−1	1.5	1.95	75.4	94.69

Because of too liberal a specification of the desirabilities, the first workpiece in Table 1 already has a desirability of 0.93, although it is a quite poor part. Figure 7 shows that the majority of the desirability index values lie between 0.8 and 1.0. Hence, between the first best part and the overall best part, the quality changes a lot.

6 Summary

In this paper, a heuristic for sequential multi-objective target value optimization was introduced. The algorithm does not need a closed form of the multivariate uncertainty distribution, but uses confidence boundaries for the single objectives. Updating the surrogate model with several points in each step, the heuristic is able to find a global optimum efficient and engineering friendly. Since every possible desirability function can be used the optimization target can be specified very flexible. However, the specification of the desirabilities has to be done very carefully, to avoid too liberal or conservative specifications. The more different confidence boundaries are used, the better the algorithm works. Due to time constraints the number of confidence boundaries should be restricted to four or five different levels. Extensive studies should still be done to give a guideline for the user, which levels give the best progress of the optimization. Furthermore, the algorithm is still rather slow when more then four objectives are considered. We are working on a variant of the heuristic, that preselects the important confidence boundaries before cross-combining them to desirability indices, to allow high-dimensional problems.

Acknowledgements Financial support for this project from the German Research Foundation (project DFG-SFB 475) is gratefully acknowledged.

References

Cox, D., & John, S. (1997). SDO: A statistical method for global optimization. In N. Alexandrow & M. Hussaini (Eds.), *Multidisciplinary design optimization: State of the art* (pp. 315–329). Philadelphia: SIAM.

Derringer, G., & Suich, R. (1980). Simultaneous optimization of several response variables. *Journal of Quality Technology, 12*, 214–219.

Forrester, A., Sóbester, A., & Keane, A. (2006). *Optimization with missing data, 462*, 935–945.
Fukuda, K., & Alain, P. (1996). Double description method revisited. In *Combinatorics and Computer Science, LNCS, 1120*, 91–111.
Harrington, J. (1965). The desirability function. *Industrial Quality Control, 21*(10), 494–498.
Henkenjohann, N., Göbel, R., Kleiner, M., & Kunert, J. (2005). An adaptive sequential procedure for efficient optimization of the sheet metal spinning process. *Quality and Reliability Engineering International, 21*, 439–455.
Henkenjohann, N., & Kunert, J. (2007). An efficient sequential optimization approach based on the multivariate expected improvement criterion. *Quality Engineering, 19*, 267–280.
Jones, D., Schonlau, M., & Welch, W. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization, 13*, 455–492.

Desirability-Based Multi-Criteria Optimisation of HVOF Spray Experiments

Gerd Kopp, Ingor Baumann, Evelina Vogli, Wolfgang Tillmann, and Claus Weihs

Abstract The reduction of the powder grain size is of key interest in the thermal spray technology to produce superfine structured cermet coatings. Due to the low specific weight and a high thermal susceptibility of such fine powders, the use of appropriate process technologies and optimised process settings are required. Experimental design and the desirability index are employed to find optimal settings of a high velocity oxygen fuel (HVOF) spraying process using fine powders ($2-8\,\mu m$). The independent factors kerosene, hydrogen, oxygen, gun velocity, stand-off distance, cooling pressure, carrier gas and disc velocity are considered in a 12-run Plackett-Burman Design, and their effects on the deposition efficiency and on the coating characteristics microhardness, porosity and roughness are estimated. Following an examination of possible 2-way interactions in a 2^{5-1} fractional-factorial design, the three most relevant factors are analysed in a central composite design. Derringer's desirability function and the desirability index are applied to find optimal factor settings with respect to the above characteristics. All analyses are carried out with the statistics software "R". The optimisation of the desirability index is done using the R-package "desiRe".

1 The Process of High Velocity Oxy-Fuel Spraying

Thermal spraying has emerged as a suitable and effective surface engineering technology to apply wear and corrosion protective coatings for various industrial applications such as tools, aero engine parts and gate valves, which are exposed to high mechanical load or intensive friction.

Among the available thermal spray processes, the High Velocity Oxy-Fuel (HVOF) spray technique is one of the most promising and preferable methods to

G. Kopp (✉)
Lehrstuhl Computergestützte Statistik, Fakultät Statistik, TU Dortmund,
Vogelpothsweg 87, 44227 Dortmund
e-mail: g.kopp@gmx.net

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_90,

produce sophisticated, wear resistant cermet coatings with exceptional high quality regarding microstructure and surface finishing. Due to high gas jet velocities and lower flame temperatures than all other thermal spray processes, HVOF spraying allows to produce coatings with extremely low porosity, low oxidizations and low carbide decomposition or carbide-matrix dissolution resulting in high hardness and high abrasion resistance (Nieminen et al. 1997).

In contrast to commonly used conventional coarse agglomerated and sintered feedstock powders with grain sizes of 15−60 μm, HVOF spraying of fine powders with grain sizes < 10 μm offers the possibility to manufacture novel superfine structured coatings with significantly improved macroscopic properties (Jia et al. 1998).

However, the feeding and processing of fine powders involves major difficulties. Due to the high surface-to-volume ratio and low specific weight, fine powders do not only provide a poor flowability in the feeding process, but also show a different thermo-kinetic behaviour during spraying. As a result, fine powders tend to overheat rapidly in the spray process leading to carbon loss of the carbide phase or the occurrence of carbide-matrix reactions, which can reduce the wear resistance of the deposited coatings. Powder agglomeration, which might encounter in the feeding or thermal spraying process represents another major problem. This effect can lead to blockages in the powder feeder system or prevent the formation of a targeted superfine structured coating morphology during the spray process.

In order to reduce such undesired effects, a sensitive selection and optimisation of the process parameter settings are necessary (Tillmann et al. 2008).

The properties of HVOF sprayed coatings depend on a number of adjustable parameters during the spray process (Turunen 2005), including the feeding parameters feeder disc velocity (FDV) and carrier gas level (CGL), the kinematic parameters stand-off distance (SOD) and gun velocity (GV) as well as the substrate temperature which is affected by the backside cooling pressure (BCP). The kerosene (KL), oxygen (OL) and hydrogen (HL) levels, which compose the combustion medium, are the main factors affecting the thermo-kinetic HVOF flame characteristic as well as the corresponding acceleration and melting behavior of the fine particles in-flight.

SOD is defined as the distance between the substrate surface and the top of the gun, where the powder injected HVOF process gas leaves the acceleration nozzle. FDV and CGL are the main factors controlling the amount of powder which is transferred into the HVOF gun. BCP indicates the cooling intensity level of the sample's backside by compressed-air convection to avoid an overheating of the substrate surface during spraying. GV controls the coating deposition process, particularly the deposition efficiency and the heat transfer to the substrate surface.

In this study fine, broken $75Cr_3C_2/25(NiCr20)$ feedstock powders with a grain size of 2−8 μm are processed to manufacture superfine structured cermet coatings with high hardness (MH), low surface roughness (RRa) and low porosity (Po) of the microstructure at high deposition efficiencies (DE).

2 Experimental Designs

In order to find optimum settings for the HVOF spraying process, three consecutive experimental designs are used, bearing the idea to gradually reduce process variables, adjust level settings and to check for significant interaction and quadratic effects.

2.1 Plackett-Burman Design

In an initial 12-run Plackett-Burman design the eight independent variables as introduced in Sect. 1 are varied at two levels according to Table 1. Assuming multiple linear models for each of the introduced response variables (Sect. 1) analysis in "R" (function lm()) yields the results summarised in Table 2. Three factors are dropped thereafter for different reasons: (1) BCP quite obviously has no effect on any of the responses and is in the following held constant at 20 psi, (2) GV has a positive effect on DE but cannot be increased any further due to machine limitations. As no interaction effects are expected GV is fixed at the highest possible setting 30,000 mm/min in further experiments, (3) CGL has a negative effect on

Table 1 Factor levels used in the experimental design stages

	Plackett-Burm.		Fractional-Fact.			Central Composite				
	−1	+1	−1	0	+1	−2	−1	0	+1	+2
KL (L/h)	7	9	8	9	10	8	9	10	11	12
HL (L/min)	60	80	60	80	100			80		
OL (L/min)	800	900	800	850	900	750	800	850	900	950
GV (mm/min)	20,000	30,000		30,000				30,000		
SOD (mm)	120	140	100	130	160	100	115	130	145	160
BCP (psi)	20	60		20				20		
CGL (L/min)	9	11		11				11		
FDV (rpm)	2.0	2.6	2.3	2.6	2.9			2.6		

Table 2 Model estimates and p-values based on the Plackett-Burman design

	MH (max.)		DE (max.)		RRa (min.)		Po (min.)	
	Est.	p-v.	Est.	p-v.	Est.	p-v.	Est.	p-v.
Int.	578.7	NA	24.92	0.0003	3.65	0.0000	1.24	0.0074
KL	127.0	NA	2.61	0.1249	−0.76	0.0001	−0.77	0.0272
HL	25.3	NA	1.18	0.4086	−0.21	0.0069	−0.17	0.4448
OL	9.3	NA	0.99	0.4827	0.43	0.0008	−0.25	0.2797
GV	−27.8	NA	4.01	0.0477	−0.03	0.4144	0.25	0.2797
SOD	−22.0	NA	−2.08	0.1905	−0.14	0.0204	0.30	0.2067
BCP	−24.5	NA	−0.71	0.6047	0.01	0.8237	0.21	0.3464
CGL	13.3	NA	−0.77	0.5793	−0.13	0.0257	−0.13	0.5333
FDV	16.7	NA	1.49	0.3155	−0.03	0.4640	0.08	0.7020

RRa, which is to be minimised. As settings beyond 11 L/min caused the formation of powder caking in the powder injection ring and the acceleration nozzle of the HVOF gun in several pilot tests, the highest reasonable level 11 L/min is used in the following designs. Despite the lack of any significant results FDV is assumed to effect in particular DE and is therefore analysed in a different range (2.3–2.9 rpm) in the following fractional factorial design. Other levels are adjusted according to Table 1. Of particular importance is the increase of both kerosene levels. Three experiments of the Plackett-Burman design run with low KL resulted in coatings too thin to allow a measurement of MH. As a consequence, no sample variance and no p-values could be estimated in the MH-model (Table 2, value NA). High KL is associated with better DE, Po and RRa in the screening model, supporting the approach to increase KL.

2.2 *Fractional-Factorial 2^{5-1} Design*

Table 3 gives selected results of the models for all four responses considering all main and 2-way interaction effects based on a fractional-factorial 2^{5-1} design with three additional center runs (level 0). None of the effect estimates left out in Table 3 are associated with p-values below 0.1. MH is mainly affected by SOD and KL, where KL should be high and SOD should be low or moderate in order to maximise MH. Figure 1a) suggests a quadratic influence of SOD regarding MH. DE is

Table 3 Selected model estimates and p-values from the fractional-factorial design

	MH (max.)		DE (max.)		RRa (min.)		Po (min.)	
	Est.	p-v.	Est.	p-v.	Est.	p-v.	Est.	p-v.
KL	55.8	0.069	7.28	0.004	0.98	0.121	−0.12	0.260
OL	1.1	0.961	−2.13	0.105	0.89	0.146	−0.01	0.912
SOD	−87.4	0.022	−7.53	0.004	−1.81	0.029	0.07	0.510
KL*OL	35.4	0.175	−1.54	0.195	0.32	0.536	−0.17	0.152
KL*SOD	19.9	0.394	−2.32	0.087	−1.75	0.031	−0.29	0.046
OL*SOD	6.7	0.761	1.54	0.196	−0.41	0.431	0.06	0.517

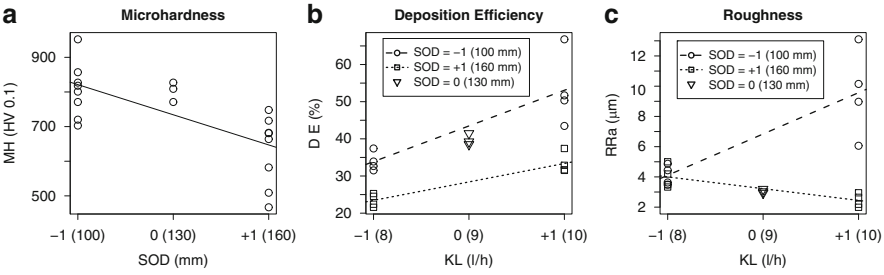


Fig. 1 Main effect and interaction plots from the fractional-factorial design

mainly determined by KL and SOD including their interaction. When holding SOD fixed, DE increases with an increase in KL. The best DE results are attained with a combination of high KL and low SOD (Fig. 1b) The same combination, however, generates very poor results for RRa (6.06–13.1 µm), while a mixture of high KL and high SOD yields the best RRa results (1.99–2.95 µm, Fig. 1c) The model fit for Po is rather poor (adj. $R^2 = 0.23$), which is most likely due to the difficult and unexact measurement procedure for porosity, which induces high variance.

2.3 Central Composite Design

For the central composite design star points are used with a distance 2 from the center to enable adjustability of KL, which can only be varied in steps of 1 L/h with the device being used. The according levels for KL, OL and SOD are captured in Table 1. Using the R function "stepAIC" (Venables and Ripley 2002) for each response the best fit is determined, allowing the algorithm to choose from all possible main and quadratic effects and all possible 2-way interactions. Starting with a model including only a constant, effects are added (or dropped) stepwise until no further improvement of the AIC criterion is possible. The resulting models are

$$MH = 784.57 + 66.77 \cdot KL + 30.90 \cdot OL^2 - 60.51 \cdot OL + 48.07 \cdot SOD^2 - 46.86 \cdot SOD + \varepsilon \qquad (se.\ 95.73, adj.\ R^2\ 0.54) \tag{1}$$

$$DE = 55.008 - 4.108 \cdot KL^2 + 5.542 \cdot KL - 7.577 \cdot OL - 4.848 \cdot SOD - 2.986 \cdot (OL \cdot KL) + \varepsilon \qquad (se.\ 6.64, adj.\ R^2\ 0.73) \tag{2}$$

$$RRa = 2.056 + 0.948 \cdot SOD^2 - 1.059 \cdot SOD + \varepsilon \quad (se.\ 1.40, adj.\ R^2\ 0.53) \tag{3}$$

For Po no effect could establish an improvement compared with the very basic model including only a constant. This is in line with the difficulties experienced in previous designs for Po, and it appears that Po cannot be properly modelled with the considered process variables. It is therefore dropped from further analysis. While residual standard errors (se.) in models (1) and (2) are considerably small, that in (3) is large, accounting for almost 3/4 of the overall mean, and results in large prediction intervals (Fig. 3c). It is, however, largely induced by one single observation made at $SOD = 100$ mm (−2), where RRa = 10.71 µm, while the mean of the remaining 17 observations is 2.44 µm and the second largest observation is 4.11 µm. Dropping this highly influencial point would result in a completely different model (using only KL^2 and KL) with an overall mean of 2.12 and residual standard error of 0.34. Results from the fractional-factorial design have shown that the observed high measurement is typical for runs at SOD = 100 mm (Fig. 1c), supporting the decision to use model (3) despite the large residual standard error.

3 Multi-criteria Optimisation

The idea of multi-criteria optimisation is to find a combination of the independent variables that optimises ideally all responses at the same time, or at least constitutes the best possible compromise.

3.1 Overlayed Contours

If the number of responses is small, as is the case here, overlayed contour-plots can assist in finding an optimum overall setting of the independent variables. Figure 2 shows the contours for MH, DE and RRa at KL levels 1 (a) and 2 (b). It is quite obvious, that MH improves while moving away in circles from (OL $=$ 1/SOD $=$ 0.5), DE gets better as OL decreases, and RRa takes its lowest value in the valley where SOD $=$ 0.56. Comparing plots (a) and (b) indicates a small advantage of KL $=$ 2 over KL $=$ 1, as e.g. in (OL $=-2$/SOD $=$ 0.56) the prediction for the most important response MH clearly exceeds 1,100 with KL $=$ 2, but with KL $=$ 1 it stays just below this mark. However, as a kerosene setting of 12 L/h (KL $=$ 2) comes with severe technical difficulties, 11 L/h (KL $=$ 1) would have to be preferred here – remembering that KL can only be varied in steps of 1 L/h (Sect. 2.3). A disadvantage of contour-plots of the original responses is that they don't answer the question of how a setting, which gives rather good results for all responses (like the one just discussed), compares with a setting that produces a bad result for one response, but an excellent one for another (such as (KL $=$ 1/OL $=-2$/SOD $=-2$)). In order to be able to make such comparisons all responses must be combined into one single measure.

3.2 Desirabilities

Before several response variables can be combined, they must be measured on a common scale. One very practical method to achieve this is to apply Derringer's

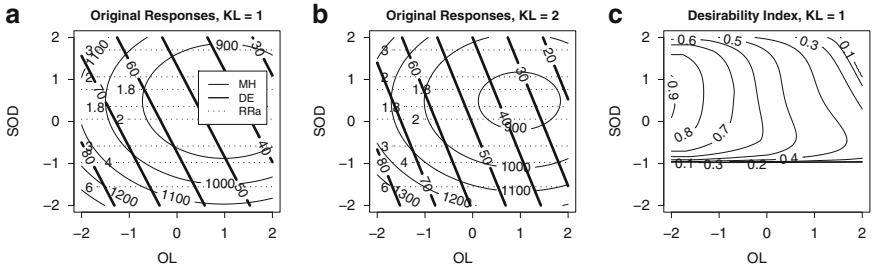


Fig. 2 Contour plots of the original responses (**a,b**) and the desirability index (**c**) using the weighted geometric mean with weights (MH $=$ 0.4/DE $=$ 0.35/RRa $=$ 0.25)

Table 4 Parameters for desirability transformations (s = 1 in all cases)

	MH (max.)	d(MH)	DE (max.)	d(DE)	RRa (min.)	d(RRa)
LSL	700	0	30	0	1	1
USL	1100	1	60	1	4	0

desirability function (Derringer and Suich 1980) (here for maximisation problems),

$$d(Y) = \begin{cases} 0 & \text{if } Y < LSL \\ (\frac{Y-LSL}{USL-LSL})^s & \text{if } LSL \leq Y \leq USL \\ 1 & \text{if } Y > USL \end{cases} \quad (4)$$

where the desirability of a response Y takes on the value 0 when Y falls below a lower specification limit (LSL), and the value 1 as Y exceeds the upper specification limit (USL). Values in between the limits are assigned desirabilities according to (4), where the transformation is linear if s = 1.

Applying the transformation to predictions made based on the models (3)–(2) yields desirabilities of these responses which can be combined into one desirability index, using e.g. the weighted geometric mean $D = \prod_{i=1}^{p} (d_i(Y_i))^{w_i}$ with weights $w_i > 0$ and $\sum_{i=1}^{p} w_i = 1$ for p quality criteria.

After transforming the original responses into desirabilities according to Table 4, using the R-package "desiRe" (Trautmann et al. 2008), contours of the desirability index with weights (MH $= 0.4/$DE $= 0.35/$RRa $= 0.25$) as shown in Fig. 2c, makes comparisons between different settings easy. For example, the two settings compared in Sect. 3.1 correspond to desirabilities of 0.92 (KL $= 1/$OL $= -2/$SOD $= 0.56$) and 0 (KL $= 1/$OL $= -2/$SOD $= -2$). This is due to unacceptable roughness predictions ($\geq$4 µm) for SOD < -1, resulting in a desirability 0.

Optimisation of the desirability index in R yields the optimum setting (KL $= 1.42/$OL $= -2.2/$SOD $= 0.56$) with a desirability of 0.93. For reasons explained in Sect. 3.1 the kerosene level 1 (11 L/h) is used in verification experiments, along with an oxygen level −2 (750 L/min) – this is preferred over -2.2 (740 L/min) as this is outside the region of experimentation – and a stand-off distance of 0.56 (138 mm).

The results of three verification experiments with this setting are captured in Fig. 3, showing the prediction lines along with their corresponding 95% prediction intervals for MH (a), DE (b) and RRa (c). All results are within the prediction limits and account for desirabilities of 0.93 (⋄), 0.83 (○) and 0.65 (□). Here, even the weakest point features quite acceptable MH (865) and RRa (1.82), its DE (61.6) achieves desirability 1. DE for the other two experiments is exceptional (76.1 and 74.4). The latter of these two also accomplishes exceptional MH (1130) with corresponding desirability 1.

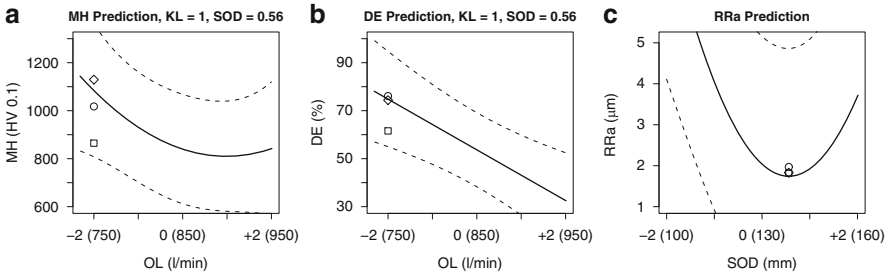


Fig. 3 Prediction line and 95% prediction interval lines (*dashed*) for MH, DE and RRa, including three verification experiments at (KL $= 1$, OL $= -2$, SOD $= 0.56$)

4 Conclusion

A HVOF spraying process was optimised for the deposition of fine, broken $75Cr_3C_2/25(NiCr20)$ feedstock powders with a grain size of $2-8\,\mu$m. After three stages of experimental design the process characteristics microhardness, deposition efficiency and roughness were transformed into desirabilities and combined into one desirability index which was optimised. Three verification experiments near the optimum gave very satisfying results with associated desirabilities between 0.65 and 0.93.

Acknowledgements The authors gratefully acknowledge the financial support of the DFG (German Science Foundation) within the Collaborative Research Centres SFB 475 and SFB 708, and the Transregional Collaborative Research Centre SFB TRR 30.

References

Derringer, G., & Suich, R. (1980). Simultaneous optimization of several response variables. *Journal of Quality Technology, 12*, 214–219.

Jia, K., Fischer, T. E., Gallois, B. (1998). Microstructure, hardness and toughness of nanostructured and conventional WC-Co composites. *Nanostructured Materials, 10*(5), 875–891.

Nieminen, R., Vuoristo, P., Niemi, K., Mäntylä, T., & Barbezat, G. (1997). Rolling contact fatigue failure mechanisms in plasma and HVOF sprayed WC-Co coatings. *Wear, 212*(1), 66–77.

Tillmann, W., Vogli, E., Baumann, I., Matthaeus, G., & Ostrowski, T. (2008). Influence of the HVOF gas composition on the thermal spraying of WC-Co submicron powders ($-8+1\,\mu$m) to produce superfine structured cermet coatings. *Journal of Thermal Spray Technology, 17*(5–6), 924–932.

Trautmann, H., Steuer, D., Mersmann, O., Ligges, U., & Weihs, C. (2008). *desiRe: Desirability functions and indices in multicriteria optimization.* R Package Version 0.9.6. From http://r-forge.r-project.org/projects/desire.

Turunen, E. (2005). *Diagnostic tools for HVOF process optimization.* PhD thesis, Finland, Espoo, VTT Publications 583.

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer. ISBN 0-387-95457-0.

Index