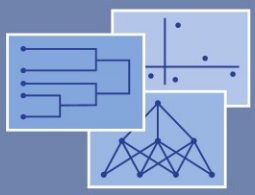
Studies in Classification, Data Analysis, and Knowledge Organization

Salvatore Ingrassia
Roberto Rocci
Maurizio Vichi *Editors*

New Perspectives in Statistical Modeling and Data Analysis



Springer

Studies in Classification, Data Analysis, and Knowledge Organization

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

Salvatore Ingrassia • Roberto Rocci
Maurizio Vichi
Editors

New Perspectives in Statistical Modeling and Data Analysis

Proceedings of the 7th Conference of the Classification and Data Analysis Group of the Italian Statistical Society, Catania, September 9-11, 2009

Springer

Editors
Prof. Salvatore Ingrassia
Università di Catania
Dipartimento Impresa
Culture e Società
Corso Italia 55
95129 Catania
Italy
s.ingrassia@unict.it

Prof. Roberto Rocci
Università di Roma "Tor Vergata"
Dipartimento SEFEMEQ
Via Columbia 2
00133 Roma
Italy
roberto.rocci@uniroma2.it

Prof. Maurizio Vichi
Università di Roma "La Sapienza"
Dipartimento di Statistica
Probabilità e Statistiche Applicate
Piazzale Aldo Moro 5
00185 Roma
Italy
maurizio.vichi@uniroma1.it

ISSN 1431-8814
ISBN 978-3-642-11362-8 e-ISBN 978-3-642-11363-5
DOI 10.1007/978-3-642-11363-5
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011930788

Cover design: deblik, Berlin

Printed on acid-free paper

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

Preface

This volume contains revised versions of selected papers presented at the seventh biannual meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, organized by the Faculty of Economics at the University of Catania in September 2009.

The conference encompassed 150 presentations organized in 3 plenary talks and 42 sessions. Moreover, one tutorial on mixture modeling took place before the meeting. With 225 attendees from 11 countries, the conference provided an attractive interdisciplinary international forum for discussion and mutual exchange of knowledge. The topics of all plenary and specialized sessions were chosen to fit the mission of CLADAG which is to promote methodological, computational and applied research within the fields of Classification, Data Analysis and Multivariate Statistics.

The chapters in this volume were selected in a second peer review process after the conference. In addition to the fundamental areas of clustering and discrimination, multidimensional data analysis and data mining, the volume contains some chapters concerning data analysis and statistical modeling in areas like evaluation, economics, finance, environmental and medical sciences, industry and services.

We would like to express our gratitude to the members of the Scientific Program committee for their ability to attract interesting contributions. We wish also thank the session organizers for inviting speakers, the chairpersons and discussants of the sessions for their stimulating comments and suggestions. We are very grateful to the referees for their careful reviews of the submitted papers and for the time spent in this professional activity. We gratefully acknowledge the Faculty of Economics of the University of Catania and the Department of Economics and Quantitative Methods, the Department of Economics and Territory, the Department of Sociology and Social Sciences, and the Faculty of Political Sciences for financial support. We are also indebted to SAS, CEUR Foundation and the Foundation for Subsidiarity for their support. A special thank is due to the Local Organizing Committee for this well-organized conference.

Finally, we would like to thank Dr. Martina Bihn of Springer-Verlag, Heidelberg, for her support and dedication to the production of this volume.

Catania *Salvatore Ingrassia*
Roma *Roberto Rocci*
Roma *Maurizio Vichi*
July 2010

Contents

Contributors

Giada Adelfio Dipartimento di Scienze Statistiche e Matematiche "S. Vianelli", Università di Palermo, Viale delle Scienze, ed. 13, 90128 Palermo (Italy), adelfio@unipa.it

Marco Alfò Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale Aldo Moro, 5, 00185 Roma, Italy, marco.alfo@uniroma1.it

Angela Alibrandi Department of Economical, Financial, Social, Environmental, Statistical and Territorial Sciences, University of Messina, Via dei Verdi 27, 98122, Messina (Italy), aalibrandi@unime.it

Patrizia Ameli Dipartimento di Istituzioni Economiche e Finanziarie, Università di Macerata, Via Crescimbeni 20–21, 62100 Macerata (Italy), patrizia.ameli@unimc.it

Mouna Akacha Department of Statistics, University of Warwick, Coventry CV4 7AL Warwick (United Kingdom), m.akacha@warwick.ac.uk

Pietro Amenta Department of Social and Economic System Analysis, Università del Sannio, Via delle Puglie 82, 82100, Benevento (Italy), amenta@unisannio.it

Graziano Aretusi D.M.Q.T.E., University of G. d'Annunzio, Viale Pindaro 42, I-65127 Pescara (Italy), graz@inwind.it

Massimo Aria Department of Mathematics and Statistics, University of Naples Federico II, Via Cinthia 26, (Monte Sant'Angelo), 80126 Napoli (Italy), aria@unina.it

Anthony C. Atkinson Department of Statistics, The London School of Economics, London WC2A 2AE (United Kingdom), a.c.atkinson@lse.ac.uk

Antonio Balzanella Dipartimento di Strategie Aziendali e Metodologie Quantitative Facoltà di Economia, Seconda Università degli Studi di Napoli, Via del Setificio, 81100 Caserta (Italy), antonio.balzanella2@gmail.com

Alessandro Barbiero Department of Economics, Business and Statistics, Università degli Studi di Milano, Via Conservatorio 7, 20122 Milano (Italy), barbiero@unimi.it

Francesco Bartolucci Dipartimento di Economia, Finanza e Statistica, Università di Perugia, Via Alessandro Pascoli 20, 06123 Perugia (Italy), bart@stat.unipg.it

Salwa Benammou Facultè de Droit & des Sciences Economiques & Politiques, Universitè de Sousse, Cité Erriadh - 4023 Sousse (Tunisia), saloua.benammou@fdseps.rnu.tn

Matilde Bini Dipartimento di Economia, Università Europea di Roma, Via degli Aldobrandeschi 190, 00163 Roma (Italy), mbini@unier.it

Giovanna Boccuzzo Dipartimento di Scienze Statistiche, Università di Padova, Via Cesare Battisti 241, 35121 Padova (Italy), boccuzzo@stat.unipd.it

Elvio C. Bonafede Department of Statistics and Applied Economics "L.Lenti", Corso Strada Nuova 65, Pavia (Italy), concetto.bonafede@unipv.it

Riccardo Borgoni Dipartimento di Statistica, Università degli Studi di Milano - Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano (Italy), riccardo.borgoni@unimib.it

Giuseppe Bove Dipartimento di Scienze dell'Educazione, Università degli Studi Roma Tre, Via del Castro Pretorio 20, 00185 Roma (Italy), bove@uniroma3.it

Rainer Brüggemann Department Ecohydrology, Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Mueggelseedamm 310. D-12587 Berlin, (Germany), brg@igb-berlin.de

Francesca Bruno Department of Statistics "Paolo Fortunati", University of Bologna, Via delle Belle Arti 41, 40126 Bologna (Italy), francesca.bruno@unibo.it

Furio Camillo Dipartimento di Statistica, Università di Bologna, Via delle Belle Arti 41, 40126 Bologna (Italy), furio.camillo@unibo.it

Maurizio Carpita Department of Quantitative Methods, University of Brescia, Contrada S. Chiara 50, 25122 Brescia (Italy), carpita@eco.unibs.it

Paola Cerchiello Department of Statistics and Applied Economics "L.Lenti", Corso Strada Nuova 65, Pavia (Italy), paola.cerchiello@unipv.it

Andrea Cerioli Dipartimento di Economia, Sezione di Statistica e Informatica, Università di Parma Via Kennedy 6, 43100 Parma Italy, andrea.cerioli@unipr.it

Marcello Chiodi Dipartimento di Scienze Statistiche e Matematiche "S. Vianelli", Università di Palermo, Viale delle Scienze, ed. 13, 90128 Palermo (Italy), chiodi@unipa.it

Daniela Cocchi Department of Statistics "Paolo Fortunati", University of Bologna, Via delle Belle Arti 41, 40126 Bologna (Italy), daniela.cocchi@unibo.it

Claudio Conversano Dipartimento di Economia, Università di Cagliari, Viale Frà Ignazio 17, I-09123 Cagliari (Italy), conversa@unica.it

Pietro Coretto Dipartimento di Scienze Economiche e Statistiche, Università di Salerno, Via Ponte Don Melillo, 84084 Fisciano (Italy), pcoretto@unisa.it

Anthony Cossari Department of Economics and Statistics, University of Calabria, Via Bucci CUBO 0C, 87036 Rende Cosenza (Italy), a.cossari@unical.it

Michele Costa Dipartimento di Scienze Statistiche "P. Fortunati", Università di Bologna, Via delle Belle Arti 41, 40126 Bologna (Italy), michele.costa@unibo.it

Giuseppina D. Costanzo Dipartimento di Economia e Statistica, Università della Calabria, Via P. Bucci, 87036 Arcavacata di Rende, Cosenza (Italy), dm.costanzo@unical.it

Valentina Cozza Dipartimento di Statistica e matematica per la ricerca economica, Università di Napoli "Parthenope", Via Medina 40 I Piano, 80133 Napoli (Italy), valecozza@unina.it

Carles M. Cuadras Department of Statistics, University of Barcelona, Diagonal 645, 08028 Barcelona (Spain), ccuadras@ub.edu

Carmen Cutugno Dipartimento di Impresa Culture e Società, Università di Catania, Corso Italia 55, 95129 Catania (Italy), carmen.cutugno@unict.it

Rosario D'Agata University of Catania, Via Vittorio Emanuele II 8, 95129 Catania (Italy), rodagata@unict.it

Luigi D'Ambra Department of Biological Sciences, Università di Napoli Federico II, Via Mezzocannone 8, 80134 Napoli (Italy), dambra@unina.it

Antonio D'Ambrosio Department of Mathematics and Statistics, University of Naples Federico II, Via Cinthia 26 (Monte S. Angelo), 80125 Napoli (Italy), antdambr@unina.it

Ida D'Attoma Dipartimento di Statistica, Università di Bologna, Via delle Belle Arti 41, 40126 Bologna (Italy), Ida.dattoma2@unibo.it

Cristina Davino Dipartimento di Studi sullo Sviluppo Economico, Università di Macerata, Via Piaggia della Torre 8, 62100 Macerata (Italy), cdavino@unimc.it

Luca De Angelis Dipartimento di Scienze Statistiche "P. Fortunati", Università di Bologna, Via delle Belle Arti 41, 40126 Bologna (Italy), l.deangelis@unibo.it

Angela De Sanctis D.M.Q.T.E., University of Pescara G. D'Annunzio, Viale Pindaro 42, I-65127, Pescara (Italy), a.desanctis@unich.it

Alfredo Del Monte Dipartimento di Economia, Università di Napoli Federico II, Via Cinthia 26 (Monte S. Angelo), 80125 Napoli (Italy), delmonte@unina.it

Laura Deldossi Dipartimento di Scienze Statistiche, Università Cattolica Del Sacro Cuore di Milano, Largo A. Gemelli 1, 20123 Milano (Italy), laura.deldossi@unicatt.it

Francesco Dell'Accio Dipartimento di Matematica, Università della Calabria, Via P. Bucci, 87036 Arcavacata di Rende, Cosenza (Italy), fdellacc@unical.it

Maria R. D'Esposito Dipartimento di Scienze Economiche e Statistiche, Università di Salerno, Via Ponte Don Melillo, 84084 Fisciano (Italy), mdesposi@unisa.it

Tonio Di Battista D.M.Q.T.E., University of Pescara G. d'Annunzio, Viale Pindaro 42, I-65127, Pescara (Italy), dibattis@unich.it

Marina Di Giacomo Dipartimento di Scienze Economiche e Finanziarie "G. Prato", Università di Torino, Corso Unione Sovietica 218 bis, 10134 Torino (Italy)

and

HERMES (Center for Research on Law and Economics of Regulated Services), Collegio Carlo Alberto, Via Real Collegio 30, 10024 Moncalieri (TO), Italy, digiacomo@econ.unito.it

Carlo Drago Dipartimento di Matematica e Statistica, Università di Napoli Federico II, Via Cinthia 26 (Monte S. Angelo), 80125 Napoli (Italy), drago@unina.it

Luigi Fabbris Dipartimento di Scienze Statistiche, Università di Padova, Via Cesare Battisti 241, 35121 Padova (Italy), luigi.fabbris@unipd.it

Marco Fattore Dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali, Università degli Studi di Milano - Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano (Italy), marco.fattore@unimib.it

Pier A. Ferrari Department of Economics, Business and Statistics, Università degli Studi di Milano, Via Conservatorio 7, 20122 Milano (Italy), pieralda.ferrari@unimi.it

Silvia Figini Department of Statistics and Applied Economics "L. Lenti", University of Pavia, Via Strada Nuova 65, 27100 Pavia (Italy), silvia.figini@unipv.it

Thaìs C.O. Fonseca Department of Statistics, University of Warwick, Coventry CV4 7AL Warwick, (United Kingdom), t.c.o.fonseca@warwick.ac.uk

Lara Fontanella D.M.Q.T.E., University of Pescara G. D'Annunzio, Viale Pindaro 42, I-65127 Pescara (Italy), lfontan@dmqte.unich.it

Antonio Forcina Dipartimento di Economia, Finanza e Statistica, Università di Perugia, Via Pascoli 10, 06100 Perugia (Italy), forcina@stat.unipg.it

Paolo Frumento Scuola Superiore S. Anna di Pisa, Piazza Martiri della Libertà 33, 56127 Pisa (Italy), p.frumento@sssup.it

Giuliano Galimberti Dipartimento di Scienze Statistiche "P. Fortunati", Università di Bologna, Via delle Belle Arti 41, 40126 Bologna (Italy), giuliano.galimberti@unibo.it

Romana Gargano Department of Economical, Financial, Social, Environmental, Statistical and Territorial Sciences, University of Messina, Via dei Verdi 27, 98122, Messina (Italy), rgargano@unime.it

Stefano A. Gattone Department SEFeMEQ, University of Tor Vergata, Rome, Via della Ricerca Scientifica 1, 00133, Roma (Italy), gattone@economia.uniroma2.it

Gianfranco Giordano Department of Mathematics and Statistics, University of Naples Federico II, Via Cinthia 26, (Monte Sant'Angelo), 80126 Napoli (Italy), gianfranco.giordano@unina.it

Giuseppe Giordano Dipartimento di Scienze Economiche e Statistiche, Università di Salerno, Via Ponte Don Melillo, 84084 Fisciano (Italy), ggiordan@unisa.it

Giorgia Giovannetti Dipartimento di Economia, Università di Firenze, Via delle Pandette 6 (FI) and European University Institute, Badia Fiesolana, Via dei Roccettini 9, I-50014 San Domenico di Fiesole (FI), Italy, giorgia.giovannetti@eui.it

Beatriz Goitisolo Department of Applied Economics III, University of Basque Country, Bilbao, Spain, Beatriz.Goitisolo@ehu.es

Jan Graffelman Universitat Politècnica de Catalunya, Department of Statistics and Operations Research, Av. Diagonal, 647, 6th floor, 08028 Barcelona, Spain, jan.graffelman@upc.edu

Luigi Grossi Università di Verona, Via dell'Artigliere 19, 37129 Verona (Italy), luigi.grossi@univr.it

Luigi Ippoliti D.M.Q.T.E., University of Pescara G. D'Annunzio, Viale Pindaro 42, I-65127 Pescara (Italy), ippoliti@unich.it

Ron S. Kenett KPA Ltd., Raanana, Israel and University of Torino, Torino (Italy), ron@kpa.co.il

Deborah Koltai Bar Ilan University, Department of Statistics, Israel, deborah.koltai@yahoo.com

Don Kristt Unit of Molecular Pathology, Laboratory of Histocompatibility and Immunogenetics, Rabin Medical Center, Petach Tikvah, Israel, pdkristt@gmail.com

Michele La Rocca Dipartimento di Scienze Economiche e Statistiche, Università di Salerno, Via Ponte Don Melillo, 84084 Fisciano (Italy), larocca@unisa.it

Fabrizio Laurini Dipartimento di Economia**,** Università di Parma, Via Kennedy 6, 43100 Parma (Italy), fabrizio.laurini@unipr.it

Yves Lechevallier INRIA, 78153 Le Chesnay cedex, (France), Yves.Lechevallier@inria.fr

Silvia Liverani Department of Mathematics, University of Bristol, (United Kingdom), s.liverani@bris.ac.uk

Marica Manisera Department of Quantitative Methods, University of Brescia, Contrada S. Chiara 50, 25122 Brescia (Italy), manisera@eco.unibs.it

Giancarlo Manzi Department of Economics, Business and Statistics, Università degli Studi di Milano, Via Conservatorio 7, 20122 Milano (Italy), giancarlo.manzi@unimi.it

Donata Marasini Dipartimento di Statistica, Università degli Studi di Milano - Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano (Italy), donata.marasini@unimib.it

Giovanni M. Marchetti Dipartimento di Statistica "G. Parenti", Università di Firenze, Viale Morgagni, 59, 50134 Firenze (Italy), giovanni.marchetti@ds.unifi.it

Paolo Mariani Dipartimento di Statistica, Università di Milano - Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano (Italy), paolo.mariani@unimib.it

Francesca Martella Dipartimento di Scienze Statistiche Sapienza Università di Roma, P.le A. Moro, 5 00185 Roma, Italia, francesca.martella@uniroma1.it

Maria C. Martini Department of Social, Cognitive and Quantitative Sciences, Università degli studi di Modena e Reggio Emilia, Viale Allegri 9, 42121 Reggio Emilia (Italy), mariacristiana.martini@unimore.it

Angelo Mazza Dipartimento di Impresa, Culture e Società, Università di Catania, Corso Italia 55, 95129 Catania (Italy), a.mazza@unict.it

Intissar Mdimagh Institut Superieur de Gestion, Universitè de Sousse, Rue Abedelaziz El Bahi - B.P. 763, 4000 Sousse, (Tunisia), mdimagh@yahoo.fr

Fabrizia Mealli Dipartimento di Statistica "G. Parenti", Università degli Studi di Firenze, Viale G.B. Morgagni 59, 50134 Firenze (Italy), mealli@ds.unifi.it

Arcangelo Merla Clinical Sciences and Bioimaging Department; Institute of Advanced Biomedical Technologies, Foundation University of Pescara G. D'Annunzio, a.merla@itab.unich.it

Simona C. Minotti Dipartimento di Statistica, Università degli Studi di Milano - Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano (Italy), simona.minotti@unimib.it

Isabella Morlini Department of Economics, University of Modena, Viale Berengario 51, 41121 Modena (Italy), imorlini@unimore.it

Mauro Mussini Dipartimento di Statistica, Università di Milano - Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano (Italy), mauro.mussini1@unimib.it

Miki Nakai Department of Social Sciences, College of Social Sciences, Ritsumeikan University, 56-I Toji-in kitamachi, Kyoto 603-8577, (Japan), mnakai@ss.ritsumei.ac.jp

Federico Niccolini Dipartimento di Istituzioni Economiche e Finanziarie, Università di Macerata, Via Crescimbeni 20 – I, 62100 Macerata (Italy), fniccolini@unimc.it

Luciano Nieddu LUSPIO, Via C. Colombo, 200, 00145 Roma, Italy, l.nieddu@luspio.it

Akinori Okada Graduate School of Management and Information Sciences, Tama University, 4-1-1 Hijirigaoka Tama city Tokyo 206-0022, okada@tama.ac.jp

Jan Owsiński Department of Systems Methods Applications, Systems Research Institute, Polish Academy of Sciences, Jan.Owsinski@ibspan.waw.pl

Barbara Pacini Dipartimento di Statistica e Matematica, Università di Pisa, Via C. Ridolfi 10, 56124 Pisa (Italy), barbara.pacini@sp.unipi.it

Francesco Palumbo Dipartimento di Scienze Relazionali "Gustavo Iacono", Università di Napoli Federico II, Via Porta di Massa 1, 80133 Napoli (Italy), fpalumbo@unina.it

Fulvia Pennoni Dipartimento di Statistica, Università degli Studi di Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano (Italy), fulvia.pennoni@unimib.it

Domenico Perrotta European Commission, Joint Research Centre, Institute for the Protection and Security of the Citizen, Via E. Fermi 1, 21020, Ispra (Italy), domenico.perrotta@ec.europa.eu

Alessandra Petrucci Dipartimento di Statistica "G. Parenti", Università di Firenze, Viale Moragni 59, 50134 Firenze (Italy), alex@ds.unifi.it

Massimiliano Piacenza Dipartimento di Scienze Economiche e Finanziarie "G. Prato", università di Torino, Corso Unione Sovietica 218 bis, 10134 Torino (Italy)

and

HERMES (Center for Research on Law and Economics of Regulated Services), Collegio Carlo Alberto, Via Real Collegio 30, 10024 Moncalieri (TO), Italy, piacenza@econ.unito.it

Marilena Pillati Dipartimento di Scienze Statistiche "P. Fortunati", Università di Bologna, Via delle Belle Arti 41, 40126 Bologna (Italy), marilena.pillati@unibo.it

Antonella Plaia Department of Statistical and Mathematical Sciences "S. Vianelli", University of Palermo, Viale delle Scienze 13, 90128, Palermo (Italy), plaia@unipa.it

Mariano Porcu Dipartimento di Ricerche Economiche e Sociali, Università di Cagliari, Viale S. Ignazio 78, 09123 Cagliari (Italy), mrporcu@unica.it

Antonio Punzo Dipartimento di Impresa, Culture e Società, Università di Catania, Corso Italia 55, 95129 Catania (Italy), antonio.punzo@unict.it

Piero Quatto Dipartimento di Statistica, Università degli Studi di Milano - Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano (Italy), piero.quatto@unimib.it

Marco Riani Dipartimento di Economia, Sezione di Statistica e Informatica, Università di Parma Via Kennedy 6, 43100 Parma Italy, mriani@unipr.it

Emilia Rocco Dipartimento di Statistica "G. Parenti", Università di Firenze, Viale Morgagni 59, 50134 Firenze (Italy), rocco@ds.unifi.it

Giulia Roli Dipartimento di Scienze Statistiche "P. Fortunati", Università di Bologna, Via Zamboni 33, 40126 Bologna (Italy), g.roli@unibo.it

Antonio A. Romano Dipartimento di Statistica e Matematica per la Ricerca economica, Università di Napoli "Parthenope", Via Medina 40 I Piano, 80133 Napoli (Italy), antonio.romano@uniparthenope.it

Elvira Romano Dipartimento di Studi Europei e Mediterranei, Seconda Università degli studi di Napoli, Via del Setificio 15, 81100 Caserta (Italy), elvira.romano@unina2.it

Mariantonietta Ruggieri Department of Statistical and Mathematical Sciences "S. Vianelli", University of Palermo, Viale delle Scienze 13, 90128, Palermo (Italy), mariantonietta.ruggieri@unipa.it

Silvia Salini Dipartimento di Scienze Economiche Aziendali e Statistiche, Università degli studi di Milano, Via Conservatorio 7, 20122 Milano (Italy), silvia.salini@unimi.it

Giuseppe Scandurra Dipartimento di Statistica e Matematica per la Ricerca Economica, Università di Napoli "Parthenope", Via Medina 40 I Piano, 80133 Napoli (Italy), giuseppe.scandurra@uniparthenope.it

Germana Scepi Dipartimento di Matematica e Statistica, Università di Napoli Federico II, Via Cinthia 26
(Monte S. Angelo), 80125 Napoli (Italy), scepi@unina.it

Manuela Scioni Dipartimento di Scienze Statistiche, Università di Padova, Via Cesare Battisti 241, 35121 Padova (Italy), scioni@stat.unipd.it

Luca Scrucca Dipartimento di Economia, Finanza e Statistica, Università degli Studi di Perugia, Via Alessandro Pascoli 20, 06123 Perugia (Italy), luca@stat.unipg.it

Biagio Simonetti Department of Social and Economic System Analysis, Università del Sannio, Via delle Puglie 82, 82100, Benevento (Italy), simonetti@unisannio.it

Gabriele Soffritti Dipartimento di Scienze Statistiche "P. Fortunati", Università di Bologna, Via delle Belle Arti 41, 40126 Bologna (Italy), gabriele.soffritti@unibo.it

Nadia Solaro Dipartimento di statistica, Università di Milano - Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano (Italy), nadia.solaro@unimib.it

Giuseppe Storti Dipartimento di Scienze Economiche e Statistiche, Università di Salerno, Via Ponte Don Melillo, 84084 Fisciano (Italy), storti@unisa.it

Isabella Sulis Dipartimento di Ricerche Economiche e Sociali, Università di Cagliari, Viale S. Ignazio 78, 09123 Cagliari (Italy), isulis@unica.it

Nicola Tedesco Dipartimento di Ricerche Economiche e Sociali, Università di Cagliari, Viale S. Ignazio 78, 09123 Cagliari (Italy), tedesco@unica.it

Venera Tomaselli University of Catania, Via Vittorio Emanuele II 8, 95129 Catania (Italy), tomavene@unict.it

Francesca Torti University of Parma, Faculty of Economics, Via J.F. Kennedy 6, 43100 Parma (Italy), francesca.torti@nemo.unipr.it

Giulio Trombetta Dipartimento di Matematica, UNICAL, Via P. Bucci, 87036 Arcavacata di Rende, Cosenza (Italy), trombetta@unical.it

Gilberto Turati Dipartimento di Scienze Economiche e Finanziarie "G. Prato", Università di Torino, Corso Unione Sovietica 218 bis, 10134 Torino (Italy)

and

HERMES (Center for Research on Law and Economics of Regulated Services), Collegio Carlo Alberto, Via Real Collegio 30, 10024 Moncalieri (TO), Italy turati@econ.unito.it

Valerio A. Tutore Department of Mathematics and Statistics, University of Naples Federico II, v.tutore@unina.it

Margherita Velucchi Dipartimento di Statistica "G. Parenti", Università di Firenze, Viale G.B. Morgagni, 59 (FI), Italy, velucchi@ds.unifi.it

Rosanna Verde Dipartimento di Studi Europei e Mediterranei, Seconda Università degli studi di Napoli, Via del Setificio 15, 81100 Caserta (Italy), rosanna.verde@unina2.it

Marika Vezzoli Department of Quantitative Methods, University of Brescia, Contrada Santa Chiara 50, 25122 Brescia (Italy), vezzoli@eco.unibs.it

Donatella Vicari Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Piazzale Aldo Moro, 5, 00185 Roma, Italy, donatella.vicari@uniroma1.it

Domenico Vistocco Dipartimento di Scienze Economiche, Università di Cassino, Via Marconi 10, 03043 Cassino (Italy), vistocco@unicas.it

Maria P. Vitale Dipartimento di Scienze Economiche e Statistiche, Università di Salerno, Via Ponte Don Melillo, 84084 Fisciano (Italy), mvitale@unisa.it

Sergio Zani Department of Economics, University of Parma, Via Kennedy 6, 43100 Parma (Italy), sergio.zani@unipr.it

Diego Zappa Dipartimento di Scienze Statistiche, Università Cattolica Del Sacro Cuore di Milano, Largo A. Gemelli 1, 20123 Milano (Italy), diego.zappa@unicatt.it

Amaya Zárraga Department of Applied Economics III, University of Basque Country, Bilbao, Spain, Amaya.Zarraga@ehu.es

Emma Zavarrone IULM, Institute of Consumer, Behaviour and Corporate Communication, Via Carlo Bo 1, 20143 Milano (Italy), emma.zavarrone@iulm.it

Part I
Data Modeling for Evaluation

Evaluating the Effects of Subsidies to Firms with Nonignorably Missing Outcomes

Fabrizia Mealli, Barbara Pacini, and Giulia Roli

Abstract In the paper, the effects of subsidies to Tuscan handicraft firms are evaluated; the study is affected by missing outcome values, which cannot be assumed missing at random. We tackle this problem within a causal inference framework. By exploiting Principal Stratification and the availability of an *instrument* for the missing mechanism, we conduct a likelihood-based analysis, proposing a set of plausible identification assumptions. Causal effects are estimated on (latent) subgroups of firms, characterized by their response behavior.

1 Introduction

The estimation of causal effects often faces problems of post-treatment complications, i.e., different sorts of selection of observations due to, e.g., nonresponse, attrition, truncation, or censoring "due to death", which may affect both observational and experimental studies (Rosenbaum 1984). In this paper, a specific endogenous selection problem is considered, namely a nonignorable missing mechanism due to nonresponse on an outcome variable. Because nonresponse occurs after treatment assignment, respondents are not comparable by treatment status: the observed and unobserved characteristics of respondents in each treatment group are likely to differ and may be associated to the values of the missing outcome, making the missing mechanism nonignorable. Often analysts use ad hoc procedures to handle missing data, such as dropping cases with missing observations, or sample mean substitution, which lead to valid inferences only under strong ignorability assumptions of the nonresponse mechanism (Little and Rubin 1987).

Within the framework of the Rubin Causal Model (RCM, Rubin 1974), a relatively recent approach to deal with post-treatment complications is Principal Stratification (PS), firstly introduced in an experimental setting (Frangakis and Rubin 2002), but with the potential of being easily extended to observational studies under specific hypotheses on the assignment mechanism. In the traditional econometric literature, the problems of endogenous selection, such as non response, are usually represented by means of selection models, whose connections with PS are

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_1,

well described by Mealli and Pacini (2008). In this paper, we use PS to address the problem of nonresponse in observational studies, where strong ignorability of the treatment holds by assumption. We further rely on the presence of an additional variable which may serve as an instrument for nonresponse. In the econometric literature, instrumental variables have already been employed to deal with nonresponse (Manski 2003). Plausible instruments for nonresponse can be relatively easily found (unlike finding instruments for other post-treatment complications), for example by data collection characteristics, which are likely to affect the response probability but not the outcome values. Characteristics of the interviewer (e.g., gender), interview mode, length and design of the questionnaire can be convincing instruments for nonresponse (see, for example, Nicoletti 2010). Here we use an instrument generating some source of exogenous variation in nonresponse in a causal inference framework. Principal strata will be defined by the nonresponse behavior in all possible combinations of treatment and instrument values. Some of the strata we will search the effect for may characterize policy relevant subpopulations, which are subgroups of the always-responders strata under the standard PS approach. We adopt a parametric perspective, using a likelihood approach, in order to achieve identification and estimation of heterogeneous causal effects on specific (latent) subgroups of units under a plausible set of identifying assumptions.

We apply our approach to the analysis of data from an observational study to evaluate the effects of subsidies to Tuscan handicraft firms on sales. In the study the percentage of nonresponse affecting the outcome ranges from 12.96% for beneficiaries to 37.62% for non-beneficiaries, and there is some evidence that nonresponse may be nonignorable. Moreover, a reasonable instrument for nonresponse is found to be the professional position of the person responding to the follow-up interview.

The paper is organized as follows. The PS framework and the likelihood approach are detailed in Sects. 2 and 3. In Sect. 4 we present the empirical analysis. Some concluding remarks are reported in Sect. 5.

2 Principal Stratification with Nonignorably Missing Outcomes and an Instrumental Variable

Let us consider a sample of N units, a vector $\mathbf{X}$ of observed pre-treatment variables, a post-treatment binary intermediate variable S, which represents nonresponse, an outcome variable of interest Y and a binary treatment T. We also consider the availability of a binary instrumental variable Z, that can be characterized with the related assumptions and regarded as an additional treatment. If unit i in the study ($i = 1, \ldots, N$) is assigned to treatment $T = t$ ($t = 1$ for treatment and $t = 0$ for no treatment) and the instrumental variable Z is regarded as an additional treatment, four potential outcomes (Rubin (1974)) can be defined for each post-treatment variable, Y and S in our case: $S_i(t, z)$, $Y_i(t, z)$ for $t = 0, 1$ and $z = 0, 1$. In observational settings, various hypotheses can be posed on the assignment mechanism. In our case, the treatment assignment for both T and Z is assumed unconfounded

Table 1 Principal strata with two binary treatments and a binary intermediate variable

G	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
S(0,0)	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
S(0,1)	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
S(1,0)	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
S(1,1)	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

given a vector X of observed pre-treatment variables:

$$T, Z \perp\!\!\!\perp S(0,0), S(0,1), S(1,0), S(1,1), Y(0,0), Y(0,1), Y(1,0), Y(1,1)|\mathbf{X}. \quad (1)$$

We further assume that in large samples for all values of $\mathbf{X}$ we can find treated and control units, as well as units with different values of the instrument (overlap assumption). In order to characterize Z as an instrument, the following exclusion-restriction assumption is imposed: $Y(0,0) = Y(0,1)$ and $Y(1,0) = Y(1,1)$, which says that for treated and for controls units, separately, the value of the instrument is unrelated to the outcome. Moreover, the instrument Z is required to have some effect on S, both under treatment and under control. Within each cell defined by specific values of the pre-treatment variables, the units under study can be stratified into 16 latent groups, named Principal Strata and denoted by $G = \{1, 2, \ldots, 16\}$, according to the potential values of $S_i(t, z)$ (see Table 1). For instance, stratum $G = 1$ includes those who would not respond under treatment and under control regardless the value of the instrument; stratum $G = 2$ includes the subgroup of units responding only under treatment and if the instrument equals 1; etc. Principal stratum membership, G_i, is not affected by treatment assignment t_i, so it only reflects characteristics of subject i, and can be regarded as a covariate, which is only partially observed in the sample (Angrist et al. 1996). When S represents a post-treatment complication, such as nonresponse, we usually need to adjust for the principal strata, which synthesize important unobservable characteristics of the subjects in the study. Note that Assumption (1) implies that

$$Y(0,0), Y(0,1), Y(1,0), Y(1,1) \perp\!\!\!\perp T, Z|S(0,0), S(0,1), S(1,0), S(1,1), X;$$

therefore, potential outcomes are independent of the treatments given the principal strata. This implication confirms the idea that, while it is in general improper to condition on $S_i(t_i, z_i)$, units can be compared conditional on a principal stratum. All the observed groups, characterized by the observed value of T, Z and S, result from a mixture of a number of principal strata. The presence of the instrument can be exploited to achieve identification of causal effects on specific latent subgroups of units. For this purpose, both parametric and nonparametric strategies can be carried out, depending on the set of assumptions that can be reasonably maintained, in order to bound and/or point-identify treatment effects. In what follows, we propose a parametric approach which introduces some distributional hypothesis, that

are consistent with the data and allow us to condition on the distribution of pre-treatment covariates. Some additional assumptions are further imposed to reduce the number of strata or stating the equivalence of the outcome distribution across strata.

3 Likelihood Approach

The likelihood-based perspective allows us to solve the identification problem, mainly thank to the results on finite mixture distribution theory (see e.g., McLachlan and Peel 2000). An additional problem, often arising in evaluation studies, is that samples are usually choice-based: program participants are oversampled relative to their frequency in the population of eligible units. Choice-based sampling is frequently adopted to reduce the costs of data collection and to get a larger number of control units. Such a design does not give rise to inferential problems if T is assumed uncounfounded and no other complications must be adjusted for. In our case this implies that we do not observe the marginal distribution of the instrument Z but the conditional distribution of $Z|T$, leading to potentially inconsistent maximum likelihood estimates (Hausman and Wise 1981). However, we can show that the conditional probability of Z, given X and T, factorizes out of the likelihood function and can be neglected in the optimization procedure, since it does not include the parameters of interest. Indeed, by letting $(T_i, Z_i, S(T_i, Z_i), Y(T_i, Z_i), \mathbf{X}_i)$ be the vector of random variables generating the observed data for each unit i, the joint probability of observing these data, under our sampling design, can be expressed as

$$\prod_{i=1}^{N} P(Z_i = z_i, S(t_i, z_i) = s_i, Y(t_i, z_i) = y_i | \mathbf{X}_i = \mathbf{x}_i, T_i = t_i).$$

It can be decomposed by considering the product of the following joint probability for each of the eight observed groups $O(t, z, s)$ defined by $T = t$, $Z = z$ and $S = s$ (where $t = 0, 1$, $z = 0, 1$ and $s = 0, 1$):

$$\prod_{i \in O(t,z,s)} P(Z_i = z, S(t_i, z_i) = s, Y(t_i, z_i) = y_i | \mathbf{X}_i = \mathbf{x}_i, T_i = t) \tag{2}$$

Equation 2 can be written as:

$$\prod_{i \in O(t,z,s)} P(S(t_i, z_i) = s, Y(t_i, z_i) = y_i | \mathbf{X}_i = \mathbf{x}_i, T_i = t, Z_i = z) \times P(Z_i = z | \mathbf{X}_i = \mathbf{x}_i, T_i = t). \tag{3}$$

The first part in (3) can be expressed as the probability of the union of events defined by the values of the potential variable $S_i(t, z)$ (i.e., by the strata membership) so that

for each observed group we have

$$\prod_{i\in O(t,z,s)}\Big[\sum_{f} P(Y(t_i,z_i)=y_i|\mathbf{X}_i=\mathbf{x}_i,T_i=t,Z_i=z,G_i=f) \times P(G_i=f|\mathbf{X}_i=\mathbf{x}_i)\Big]\times P(Z_i=z|\mathbf{X}_i=\mathbf{x}_i,T_i=t)$$

where the index f denotes the latent strata memberships G_i of units in the observed group $O(t,z,s)$. $P(Z_i=z|\mathbf{X}_i=\mathbf{x}_i,T_i=t)$ can be omitted since it is ancillary with respect to the parameters of interest characterizing the distribution of Y and the strata proportions. In addition note that, due to Assumption (1),

$$P(Y(t_i,z_i)=y_i|\mathbf{X}_i=\mathbf{x}_i,T_i=t,Z_i=z,G_i=f)=P(Y(t_i,z_i)=y_i|\mathbf{X}_i=\mathbf{x}_i,G_i=f).$$

Within the PS framework and under the assumption of unconfoundedness, the likelihood function is shown to result in a finite mixture of distributions and identification is straightforward. Indeed, by exploiting the correspondence between the observed groups $O(t,z,s)$ and the latent strata, we obtain the likelihood function to be maximized where, without further assumptions, all the 16 latent strata are involved:

$$\begin{aligned}
&\prod_{i\in O(1,1,1)}\sum_{k} P(Y(t_i,z_i)=y_i|\mathbf{X}_i=\mathbf{x}_i,G_i=k)\times P(G_i=k|\mathbf{X}_i=\mathbf{x}_i)\\
&\times\prod_{i\in O(1,1,0)}\sum_{j} P(G_i=j|\mathbf{X}_i=\mathbf{x}_i)\\
&\times\prod_{i\in O(1,0,1)}\sum_{h} P(Y(t_i,z_i)=y_i|\mathbf{X}_i=\mathbf{x}_i,G_i=h)\times P(G_i=h|\mathbf{X}_i=\mathbf{x}_i)\\
&\times\prod_{i\in O(1,0,0)}\sum_{l} P(G_i=l|\mathbf{X}_i=\mathbf{x}_i)\\
&\times\prod_{i\in O(0,1,1)}\sum_{m} P(Y(t_i,z_i)=y_i|\mathbf{X}_i=\mathbf{x}_i,G_i=m)\times P(G_i=m|\mathbf{X}_i=\mathbf{x}_i)\\
&\times\prod_{i\in O(0,1,0)}\sum_{o} P(G_i=o|\mathbf{X}_i=\mathbf{x}_i)\\
&\times\prod_{i\in O(0,0,1)}\sum_{p} P(Y(t_i,z_i)=y_i|\mathbf{X}_i=\mathbf{x}_i,G_i=p)\times P(G_i=p|\mathbf{X}_i=\mathbf{x}_i)\\
&\times\prod_{i\in O(0,0,0)}\sum_{q} P(G_i=q|\mathbf{X}_i=\mathbf{x}_i) \qquad (4)
\end{aligned}$$

where k, j, h, l, m, n, o, p and q index the different latent strata included in each observed group (e.g., $k=2,4,6,8,10,12,14,16$).

In order to form the likelihood function, we have to specify a distribution for the potential outcomes conditional on the observed pre-treatment covariates, as well as a model for the principal strata membership G. The likelihood function results in a finite mixture model. Thank to the exclusion-restriction assumption some of such parameters are constrained to be the same. Moreover, in order to reduce the number of strata, different monotonicity assumptions can be imposed, which can relate both to the response behavior w.r.t. the instrument and to the response behavior w.r.t. the treatment. Note that monotonicity assumptions cannot be falsified by data, but only reasonably assumed. Once parameter estimates are obtained, both the strata proportions and the causal effects of interest can be estimated, averaging over the observed distribution of covariates, as a function of such parameters.

4 Empirical Study

In the framework of the Programs for the Development of Crafts in Tuscany, Italy (Regional Law n. 36, 4/4/95), we evaluate the effects of credit on investments (PSA 2003–2005), provided in years 2003 and 2004 by the Regional government under the form of soft loans, on sales in year 2005. Data are obtained by integrating different data sources: administrative archives provided by the Chamber of Commerce (2001–2004) and by the Internal Revenue Service (2002), data on assisted firms collected by the Regional Government, and data from a telephone survey (Mattei and Mauro 2007). The survey was conducted in order to gather additional information, not contained or not yet available in administrative archives, in particular 2005 outcome variables of firms' performances (sales, number of employees, production innovation). We consider a sample of 108 beneficiaries and of 707 non-beneficiaries firms. The binary treatment T is thus defined as an indicator of whether a firm receives a public soft loan. Previous evidence (Mattei and Mauro 2007) shows that nonresponse on sales may be nonignorable. In order to estimate the treatment effect on sales (2005), Y, in the presence of nonresponse (denoted by S), we use an indicator variable, Z, which assumes value 1 if the owner responds to the phone interview and 0 if an employee responds, as an instrument for nonresponse. Z can be thought of as an additional intervention, and as a reasonable instrument for nonresponse, because (a) the propensity to provide information on sales may vary and may depend on the person job task, with owners being a priori more inclined to answer questions on sales; (b) we can reasonably assume that, conditional on observed firm characteristics, the person responding to the interview is determined by random events, such as time of the interview, absence/presence in the office, etc., not directly related to the outcome variable. Indeed, the percentages of cases where owner responds to the interview in both treatment arms are very close (73.15% for treated and 77.93% for control units). We maintain the following assumptions: unconfoundedness of T and Z and overlap assumptions, exclusion-restriction, lognormality of sales Y

conditional on the principal strata and on the vector of pre-treatment variables,[1] and multinomial logit for the distribution of principal strata. All the hypotheses are assumed to be valid conditional on the vector X of observed pre-treatment variables. The following monotonicity assumption[2] is also imposed:

$$S(t,0) \leq S(t,1) \quad \forall t \tag{5}$$

$$S(0,z) \leq S(1,z) \quad \forall z. \tag{6}$$

Assumption (5) relates to the response behavior w.r.t. the instrument: for a fixed treatment level, units responding when an employee answers to the phone interview (i.e. $Z = 0$) would respond also when the owner answers ($Z = 1$). Analogously, Assumption (6) relates to the response behavior w.r.t. the treatment: for a fixed value of the instrument, units responding under control would respond also when treated. In this application, both monotonicity assumptions seem to be plausible, because we may reasonably assume that exposure to the treatment, i.e., the receipt of public incentives, makes the interviewed person more responsive to administrative requests, and also that owners may be more willing to take the responsibility to declare the amount of sales. Assumptions (5) and (6) imply the non-existence of some of the 16 strata in Table 1. In particular, the number of strata are reduced to be 6 (i.e., strata 1, 2, 4, 6, 8 and 16). Note that, in this case, the strata containing information on causal effects, in the sense that some units respond under treatment and some other units respond under control, are strata 6, 8 and 16, so that the goal is to identify and estimate causal effects on (union of) strata 6, 8 and 16. To simplify the model, we constrained the slope coefficients and the variances to be the same in all the strata, but separately under treatment and under control. The exclusion-restriction assumption further reduces the number of parameters to be estimated by equating the outcome distribution parameters of groups for which we observe $T = 1; Z = 0$ and $T = 1; Z = 1$ or, analogously, $T = 0; Z = 0$ and $T = 0; Z = 1$. Under this set of assumptions the estimated proportions of strata 4 and 8 are not significantly different from 0. Therefore, we can suppose these strata do not exist and hope to learn something about the causal effect $E[Y(1) - Y(0)|G = g]$ only for strata 6 and 16. Specifically, stratum 6 includes units who would respond only if owner answers to the phone interview. Stratum 16, which results to be to largest one, includes units responding under treatment and under control regardless of the value of the instrument. Results are reported in Table 2 and show different, positive, but negligible effects of the incentives in the two subgroups of firms. Moreover, the average treatment effect for $G \in \{6, 16\}$ is 227, 658.1 euros, which is smaller than the one estimated by neglecting the nonignorability of nonresponse, improperly comparing the 94 respondents under treatment with the 441 under control (449, 524.8 euros,

[1] Pre-treatment characteristics considered in the present analysis are: number of employees, sector of activity and taxable income in the year 2002.

[2] Several sets of monotonicity assumptions have been compared, by computing the scaled log-likelihood ratio, in the spirit of a direct likelihood approach.

Table 2 Estimation results (standard errors are computed using the delta method)

G	$\widehat{P}(G=g)$	(se)	$\widehat{E}[Y(1)-Y(0)\|G=g]$	(se)
1	0.209	(0.031)		
2	0.140	(0.033)		
6	0.064	(0.013)	37,634.6	(52,910.4)
16	0.586	(0.020)	248,526.0	(210,310.4)
(6,16)			227,658.1	(190,185.0)

s.e. $= 112,626.5$) and quite similar to the one estimated by a standard PS approach for the larger group of always responders ($275,073.5$ euros, s.e. $= 127,691.2$).

5 Concluding Remarks

We considered the problem of nonignorable nonresponse on an outcome in a causal inference framework and use PS as a tool to represent post-treatment complications within the RCM. By exploiting a binary instrumental variable for nonresponse and by adopting a likelihood approach, we identified and estimated causal effects on some latent subgroups of units. We showed that our strategy also allowed us to take account of choice-based sampling. The parametric approach rests on distributional assumptions that are consistent with the data. Note that when specifying a mixture model, parametric assumptions refer to distributions of the outcome variables conditional on the strata, while the underlying distribution of unobservables determining the principal strata is not parametrically specified. In this context, parametric assumptions have explicit implications on the probability law of variables within observed groups, so that the theory of mixture models can be exploited for both identification and specification testing (Mealli and Pacini 2008).

References

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association*, *91*, 444–472.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, *58*, 191–199.

Hausman, J. A., & Wise, D. (1981). Stratification on endogenous variables and estimation: The gary income maintenance experiment. In C. Manski & D. McFadden (Eds.), *Structural analysis of discrete data with econometric applications*. Cambridge, MA: MIT Press.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley.

Manski, C. F. (2003). *Partial identification of probability distributions*. New York: Springer-Verlag.

Mattei, A., & Mauro, V. (2007). *Evaluation of policies for handicraft firms*. Research Report, www.irpet.it/storage/pubblicazioneallegato/153_Rapporto%20artigianato.pdf

McLachlan, G., & Peel, D. (2000). Finite mixture models. Wiley, New York.
Mealli, F., & Pacini B. (2008). Comparing principal stratification and selection models in parametric causal inference with nonignorable missingness. *Computational Statistics and Data Analysis*, *53*, 507–516.
Nicoletti, C. (2010). Poverty analysis with missing data: Alternative estimators compared. *Empirical Economics*, 38 (1), 1–22.
Rosenbaum, P. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A*, *147*, 656–666.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized Studies. *Journal of the Educational Psychology*, *66*, 688–701.
Rubin, D. B. (2006). Causal inference through potential outcomes and principal stratification: Application to studies with censoring due to death. *Statistical Science*, *21*, 299–321.

Evaluating Lecturer's Capability Over Time. Some Evidence from Surveys on University Course Quality

Isabella Sulis, Mariano Porcu, and Nicola Tedesco

Abstract The attention towards the evaluation of the Italian university system prompted to an increasing interest in collecting and analyzing longitudinal data on students' assessments of courses, degree programs and faculties. This study focuses on students' opinions gathered in three contiguous academic years. The main aim is to test a suitable method to evaluate lecturer's performance over time considering students' assessments on several features of the *lecturer's capabilities*. The use of the same measurement instrument allows us to shed some light on changes that occur over time and to attribute them to specific characteristics. Multilevel analysis is combined with Item Response Theory in order to build up specific trajectories of performance of *lecturer's capability*. The result is a random-effects ordinal regression model for four-level data that assumes an ordinal logistic regression function. It allows us to take into account several factors which may influence the variability in the assessed quality over time.

1 Introduction

In Italy, the assessment of teaching quality in students' perception is a mandatory task for each public university institution. This study aims to build up an overall measure of *lecturer's capability* considering the evaluations she/he received in three academic years (a.y.) from her/his students who may have attended different courses held by the lecturer in the three a.y.; thus *lecturer's capability* in teaching is measured by considering evaluations gathered in different classes and which may concern different courses. A multilevel Graded Response Model (Adams 1997, Grilli and Rampichini 2003, Sulis 2007) is adopted to evaluate lecturers' performances over time considering students' responses to a set of selected items of the questionnaire for the evaluation of teaching activities. In this framework the selected items are supposed to have the same discrimination power. It is important to underline that students who evaluate the same lecturer change over years, whereas the lecturer who is the *object* of the evaluation does not. Hence, we are not moving in the classical framework of longitudinal analysis where repeated measurements on

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_2,

the same students are observed at each time t. The study moves from the perspective that the evaluation of lecturers' capability over time allows us to take strictly into account of both the multivariate structure of the responses provided by students and the characteristics that vary over time.

The modeling approach here adopted lies in the framework of Generalized Linear Mixed Models (Hedeker and Gibbons 1994, Gibbons and Hedecker 1997, Agresti et al. 2000, Hedeker et al. 2006): specifically, it can be set as a four-level random-effects regression model assuming an ordinal logistic regression function. This model allows us to describe relationships across lecturer's evaluations over years taking into account possible sources of heterogeneity which may occur across units at different hierarchical levels. However, in this study, it is not considered the heterogeneity which may occur across evaluations gathered in different courses taught by a lecturer in the same a.y.. The recent changes in the Italian university system required several adjustments in the denomination of university courses and in the reorganization of the degree programs; this makes hardly possible to analyze lecturer's evaluations over time by considering just evaluations on the lecturer gathered from the same course in the three a.y.. The main purpose of this work is to make an attempt to overcome the effect of seasonal/annual disturbances which can alterate students' perception of *lecturer's capability* with the aim to provide an overall measure of performance. However, a discussion is attempted on the further potentialities of the approach as a method to build up *adjusted* indicators of *lecturer's capability* in which the effects of factors which make evaluations not comparable are removed.

2 The Data

The data used in this application are provided by the annual survey carried out at the University of Cagliari to collect students' evaluations on the perceived quality of teaching. The analysis concerns questionnaires gathered at the Faculty of Political Sciences. Three different waves have been considered, namely those carried out at 2004/05, 2005/06 and 2006/07 a.y. Students' evaluations are collected by a questionnaire with multi-item Likert type four-categories scales. A bunch of items addressed to account for specific features of the lecturers' capabilities have been selected: $I_1 =$ *prompt student's interest in lecture*, $I_2 =$ *stress relevant features of the lecture*, $I_3 =$ *be available for explanation*, $I_4 =$ *clarify lecture aims*; $I_5 =$ *clearly introduce lecture topics*; $I_6 =$ *provide useful lectures*.

A number of 47 lecturers have been considered in the analysis; specifically: those who received at least 15 evaluations per a.y. (the total number of evaluations per lecturer ranges from 15 up to 443). In the three a.y., 10,486 evaluation forms have been gathered: 3,652 in the first a.y., 3,123 in the second and 3,711 in the third. According to the academic position, the 47 lecturers are divided in four categories: 17 full professors, 15 associated professors, 13 researchers and 2 contractors. The subject areas are seven: law (8 lecturers), economics (9), geography (2), foreign languages (2), sociology (7), mathematics and statistics (6), history and political sciences (13).

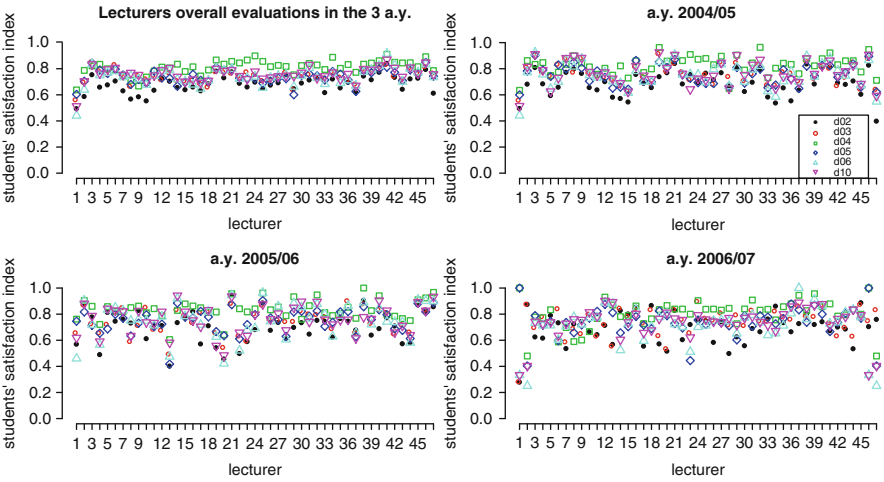


Fig. 1 z' indexes

The z' index to measure the dissimilarity among categorical ordered distributions has been used in order to summarize the evaluations concerning the same lecturer in the three a.y.. Each z'_i compares the observed cumulative distribution of students' responses (F_{I_i}) to the K-categories of item I_i, with the hypothetical cumulative distribution ($F_{I_{i.LS}}$) that we would have observed if all the evaluators would have scored for item I_i the category which marks the *lowest intensity of satisfaction* (Capursi and Porcu 2001)

$$z'_i = \frac{1}{K-1} \sum_{k=1}^{K-1} |F_{I_{i.k}} - F_{I_{i.LS.k}}|.$$

The four graphs in Fig. 1 display the values of the z' index by item and lecturers and by item, lecturers, and a.y.. The index which summarizes the evaluations in the three a.y. shows lower variability (sd ranges from 0.0535 to 0.0721) than distributions of z' in each a.y. (sd_{2004}: 0.0725 ÷ 0.0979, sd_{2005}: 0.0778 ÷ 0.1313, sd_{2006}: 0.1109 ÷ 0.1567). This suggests that more information on lecturer's evaluations could be gathered from an analysis which considers lectures' capability over time (Giusti and Varriale 2008).

3 Modeling Lecturers' Capability Over Time

The modeling approach frequently adopted to cope with multi-items Likert type indicators arises from the Item Response Theory (IRT). In this framework items are indicator variables of an unobservable latent trait. Grilli and Rampichini (2003) show as the data structure with multiple items can be simplified by

re-parameterizing the multivariate model with I items as a two-level univariate model with a dummy bottom variable. The new parametrization deals with multivariate indicators using a multilevel modeling approach where subject j (for $j = 1, \ldots, n$) denotes *level*-2 unit and where there are I (for $i = 1, \ldots, I$) responses (*level*-1) from the new bottom level. In this way the multivariate regression model is handled as univariate and standard routines for multilevel model can be used. In multilevel models the correlation brought about sources of heterogeneity is taken into account at any level by including random-effects at each stage of the hierarchy. Hedeker and Gibbons (1994) develop an ordinal random-effects regression model suitable to deal with either longitudinal and clustered observations; furthermore they show (Gibbons and Hedecker 1997) as the *two-level* model, frequently applied in the literature for modeling clustered units, longitudinal studies or repeated observations on the same subject, can be extended to deal with combination of these structures (e.g. clustered and repeated) by setting a *three-level* model: e.g. repeated responses of the same students clustered in questionnaires (*level*-2 units) which are furthermore nested in courses (*level*-3 units) (Grilli and Rampichini 2003, Sulis 2007); respondents grouped in parliamentary constituencies which belong to surveys held in three different years (Agresti et al. 2000).

It is a common practice to formulate the IRT models as two-level hierarchical models where *level-1* units are responses to each item whereas *level-2* units are questionnaire which evaluate *lecturer's capability* (Van den Noortgate and Paek 2004). Indicating with n the number of students and with I the number of items, the total number of level-1 observations is $n \times I$. IRT models for ordered responses assumes that for each item in the questionnaire the probability to observe a specific response category relies on the threshold or cut-point parameters (or *item parameters*) which characterize the categories of the items and on a *subject parameter*, also called (in the psychometric field) *ability parameter* (Rasch 1960). The former are interpreted as a kind of *difficulty parameters* since they signal how much difficult is to cross a category of a specific item. The ability parameters are individual estimates of the unobservable latent trait *lecturer's capability* in students' perception (Sulis 2007). An additional parameter (*discrimination parameter*) could be considered whenever items in the questionnaire are supposed to have different discrimination power. Combining the multilevel and the IRT framework, *person parameters* are the random intercepts which characterize responses arisen from the same questionnaire. The ability parameters on which this study focuses on are the *lecturer's overall capability* in the three a.y. and their variability over time. The questionnaires (which are *level*-2 units) are clustered according both to which a.y. the survey has taken place and to which lecturer the evaluation is addressed. Hence, the parameters considered are random terms which account for correlations across evaluations of the same lecturer and across the three a.y.

3.1 A Four-Level Ordinal Logistic Mixed-Effects Model

Let $\boldsymbol{Y}_{jtl}$ be the vector pattern of ordinal item responses of subject j which evaluates lecturer l in the t a.y. The ordered categories ($k = 1, \ldots, K$) of item can

be considered as values of an underlying continuous variable Y^*_{ijtl} ($Y^*_{ijtl} = k$ if $\tau_{i(k-1)} \leq Y_{ijtl} \leq \tau_{ik}$) which is supposed to have a logistic distribution. Denoting with η a cumulative ordinal logistic link function (Gibbons and Hedecker 1997)

$$\eta_{i_{(k)}jtl} = logit[P(Y_{ijtl} \leq k)] = \tau_{ik} - (\lambda_{jtl} + \theta_{tl} + \zeta_l) \tag{1}$$

$\lambda_{jtl} \sim N(0, \sigma^2_\lambda)$, $\theta_{tl} \sim N(0, \sigma^2_\theta)$ and $\zeta_l \sim N(0, \sigma^2_\zeta)$ are three random terms which account for unobserved heterogeneity at different levels of the hierarchy. Each of $K-1$ *logit* expresses the ratio between the probability to score category k or lower of item i evaluating lecturer l in the t a.y. on the probability to score higher categories as function of (a) a threshold item parameter (τ_{ik}), (b) a student parameter λ_{jtl} and (c) two lecturer parameters θ_{tl} and ζ_l. The items in the evaluation forms are supposed to have the same power to discriminate across lecturers and students with different intensity of the latent trait. To sum up, the model has a hierarchical structure with four levels: (i) item responses are *level*-1 units; (ii) evaluation forms are *level*-2; (iii) lecturers' evaluation forms by year combination are *level*-3 units (iv) lecturers' evaluation forms in the three a.y. are *level*-4 units. The *level*-4 random effect ζ_l (for $l = 1, \ldots, L$) is considered the lecturer's parameter which is shared by evaluations addressed to the same lecturer in the three a.y.; the *level*-3 random parameter θ_{tl} accounts for year-to-year variation in log-odds ratio for evaluations of the same lecturer (for $t = 1, 2, 3$); *level*-2 random parameter λ_{jtl} is the student's parameter which accounts for correlations between responses on the same student (variability between responses in the same evaluation form). Namely, the model allows the *level*-3 random intercept (θ_{tl}) to vary randomly around the mean of a generic *level*-4 random intercept (ζ_l) which accounts for "lecturer l overall capability" (Agresti et al. 2000, Adams 1997).

Model 1 assumes that the random effect θ_{tl} has a normal distribution (rather than a tri-variate normal) but it introduces a further random term (ζ_l) to take into account of the intra-class correlation which may occur across evaluations of the same lecturer gathered in the three a.y.. This parametrization with θ_{tl} univariate implicitly constrains to be equal the variance between questionnaires which evaluate the same lecturer in each of the three a.y. and the correlations between pairs of years (Agresti et al. 2000). Thus, adding up an additional level in the hierarchy structure leads to a more parsimonious model in terms of number of parameters: in Model 1 the number of fixed effects are $I \times (K-1)$ threshold parameters and the three unknown variances of the random terms (σ^2_λ, σ^2_θ and σ^2_ζ), whereas in the *level*-3 model with θ_{tl} tri-variate normal the parameters of the random part of the model are 7 (σ^2_λ, $\sigma^2_{\theta_1}$, $\sigma^2_{\theta_2}$, $\sigma^2_{\theta_3}$, $\sigma_{\theta_1,\theta_2}$, $\sigma_{\theta_1,\theta_3}$, $\sigma_{\theta_2,\theta_3}$).

Comparisons across threshold parameters of different items express the *difficulty* of different facets of the teaching. These parameters allow to highlight those aspects of teaching (measured throughout specific items) which require a higher or lower *lecturer's capability* in order to gain a positive assessment. Moreover, the greater *lecturer's capability* is the higher the probability to receive in each item an excellent evaluation. The means of the posterior distributions of the three random terms, obtained by using *empirical bayes estimates*, can be interpreted as estimates

Table 1 Four-level multilevel model

Item	τ_{i1}	se(τ_{i1})	τ_{i2}	se(τ_{i2})	τ_{i3}	se(τ_{i3})
I_1 to prompt student's interest in lecture	−4.470	(.082)	−2.343	(.072)	.925	(.070)
I_2 to stress relevant future of the lecture	−5.511	(.092	−3.199	(.074)	.130	(.070)
I_3 to be available for further explanation	−6.252	(.107)	−4.346	(.081)	−.869	(.070)
I_4 to clarify lectures aims	−5.511	(.093)	−3.137	(.074)	.112	(.070)
I_5 to clearly introduce lecture topics	−5.115	(.087)	−3.105	(.074)	.067	(.070)
I_6 to provide useful lectures	−5.432	(.090)	−3.360	(.075)	−.234	(.070)
Random effects	$var(\zeta_l)$	se[var(ζ_l)]	$var(\theta_{tl})$	se[var(θ_{tl})]	$var(\lambda_{jtl})$	se[var(λ_{jtl})]
	.809	(.0642)	.485	(.049)	4.747	(.089)

Statistical Software: GLAMM (Rabe-Hesketh et al. 2004), log likelihood −51,433.203.
Maximization method adopted: marginal maximum likelihood with Gauss-Hermite quadrature.

of the three latent variables (Sulis 2007): student's perceived quality of lecturer's capability, variability in lecturer's capability in the three a.y., and lecturer's overall capability. The corresponding posterior standard deviations are often interpreted as standard errors in the IRT framework.

Results of Model 1 are depicted in Table 1. The Intra-class Correlation Coefficient (ICC) shows how the unexplained variability in the latent responses is distributed across levels; thus it is a measure of how much high is the similarity across units which belong to the same cluster. In Model 1, where the latent variable is specified to follow a logistic distribution, the within *level*-1 variability is set equal to $\pi^2/3$. The estimates of σ^2_λ, σ^2_θ and σ^2_ζ (Table 1) provide information on the amount of unexplained variability at each level. Specifically, as it could be expected, about 51% of the variability in the responses is explained by the fact that *level*-2 units cluster repeated measurements on the same student (who evaluates several features of teaching). Thus this source of the heterogeneity is the result of the different perception that students have of teaching quality. The remaining 14% of the unexplained variability is ascribable to the variability in the assessments observed across evaluation forms addressed to different lecturers. The variability between lecturer's evaluations is a combination of the two random effects θ_{tl} and ζ_l. The fraction of variability, even though it is ascribable to lectures' performances, can be further decomposed into two parts: (i) a fraction of heterogeneity in the data which is given to unobservable specific characteristics/qualities concerning lecturers' capability and invariant across the three a.y. and (ii) a fraction which capture unobservable factors which may vary. The former is described by the variance of the random term ζ_l (e.g. *lecturer's capability* of teaching) and accounts for 9% of the variability in the evaluations; the latter is described by the variance of the random term θ_{tl} and explains about the 5%. Hence, the variance ratio between the evaluations of the same lecturer in two different a.y. is equal to 0.625 (Agresti et al. 2000). The source of variability reproduced by θ_{tl} can be ascribed to several factors which can rise heterogeneity in the data and which are not observed in this framework

Table 2 Posterior estimates of students and lecturers parameters: some descriptive statistics

Statistics	$\hat{\lambda}_{jtl}$	$\hat{\theta}_{04.l}$	$\hat{\theta}_{05.l}$	$\hat{\theta}_{06.l}$	$\hat{\zeta}_l$
Min.	−8.36	−1.40	−1.15	−1.25	−2.37
1st Qu.	−1.21	−0.36	−0.17	−0.42	−0.35
Median	0.09	0.05	0.15	−0.09	0.13
Mean	0.03	0.07	0.11	−0.09	0.11
3rd Qu.	1.51	0.41	0.46	0.31	0.46
Max.	4.76	1.24	1.03	1.25	2.45

(e.g. the different background of the students who evaluate; the total workload of the lecturer; the specific topic of the course; the number of students in the classroom; etc.). Descriptive statistics related to the posterior estimates of the three latent variables are depicted in Table 2. The posterior estimates $\hat{\zeta}_l$ of *lecturers' capability* allows to make comparisons across lecturers.

Looking at the values of the cut-points of the categories and their standard errors (Table 1), it is interesting to see where they are located in the continuum which here represents the two latent variables: "students perception of teaching quality" and "teacher's overall capability". The easiest task of teaching for a lecturer seems being *available for explanations* (I_3). The level of students' satisfaction required to observe the highest positive response *definitely yes* ($\tau_{33} = -0.869$) is located well below the average and the median value. Furthermore, the values of the quantiles of the distribution of $\hat{\zeta}_l$ indicate that a relevant rate of lectures have in average a *definitely positive* score in the item. At the other end of the continuum there is the item *prompt student's interest in lecture* (I_1). To cross the first cut point of this item (e.g. to be more satisfied than unsatisfied) is almost as difficult as to cross the second cut point of item I_3. The cut points of items I_2, I_4, I_5, I_6 are close and the difference across them in terms of intensity are not statistically significant. This means that it is required almost the same level of teacher's capability and students' satisfaction in order to cross the categories of the four items.

Model 1 is a descriptive model (Wilson and De Boeck 2004) since it considers just random intercepts ignoring the effect of items, students, or lecturer by year (*level*-1, *level*-2, *level*-3) covariates. These factors could be specifically taken into account in the analysis by introducing *level*-2 -*x*- , or *level*-3 -*z*- or *level*-4 *u* covariates – depending whether or not we are dealing with time-dependent or time-independent covariates – which may affect lecturer's capability (Adams 1997, Zwiderman 1997, Sulis 2007). The effect of covariates can be specified in different ways: by allowing covariates to affect directly the ability parameters or indirectly the responses (Sulis 2007). For instance, if time dependent variables are supposed to influence lecturer's capability

$$\theta_{tl} = \sum_{s=1}^{S} \gamma_s z_{lst} + \varepsilon_{lt} \quad and \quad \zeta_l = \sum_{c=1}^{C} \alpha_c u_{lc} + \epsilon_l.$$

The model with covariates allows to sort out *adjusted* estimates of lecturers' capability parameters. This means that it makes possible comparisons across lecturers controlling for those factors as e.g. the lecturer's experience, the topic thought by the lecturer, the number of students in the class, etc., which make lecturers' evaluations not comparable. Furthermore, the heterogeneity across evaluations of the same lecturer gathered in different courses may be partially controlled by considering in the model a specific covariate which takes into account for the year of enrollment of students. The low number of lecturers observed in this application (47) did not allow to pursue this specific task.

References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistiscs*, *22*, 47–76.

Agresti, A., Booth, G., Hobert, O., & Caffo, B. (2000). Random-effects modeling of categorical response data. *Sociological Methodology*, *30*, 27–80.

Capursi, V., & Porcu, M. (2001). La didattica universitaria valutata dagli studenti: un indicatore basato su misure di distanza fra distribuzioni di giudizi. In *Atti Convegno Intermedio della Societ Italiana di Statistica 'Processi e Metodi Statistici di Valutazione', Roma 4–6 giugno 2001*. Società Italiana di Statistica.

Gibbons, R. D., & Hedecker, D. (1997). Random effects probit and logistic regression models for three-level data. *Biometrics*, *53*(4), 1527–1537.

Giusti, C., & Varriale, R. (2008). chapter Un modello multilivello per l'analisi longitudinale della valutazione della didattica. In *Metodi, modelli e tecnologie dell'informazione a supporto delle decisioni* (Vol. 2, pp. 122–129). Franco Angeli, Pubblicazioni DASES.

Grilli, L., & Rampichini, C. (2003). Alternative specifications of multivariate multilevel probit ordinal response models. *Journal of Educational and Behavioral Statistics*, *38*, 31–44.

Hedeker, D., Berbaum, M., & Mermelstein, R. (2006). Location-scale models for multilevel ordinal data: Between- and within-subjects variance modeling. *Journal of Probability and Statistical Sciences*, *4*(1), 1–20.

Hedeker, D., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, *50*(4), 993–944.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Gllamm manual. *U. C. Berkeley Devision of Biostatistics Working Paper Series*, 160.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: Mesa Press.

Sulis, I. (2007). *Measuring students' assessments of 'university course quality' using mixed-effects models*. PhD thesis, Università degli Studi di Palermo, Palermo.

Van den Noortgate, W., & Paek, I. (2004). Person regression models. In *Explanatory item response models: A generalized linear and non linear approach* (pp. 167–187). New York: Springer.

Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In *Explanatory item response models: A generalized linear and non linear approach* (pp. 43–74). New York: Springer.

Zwiderman, A. H. (1997). A generalized rasch model for manifest predictors. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*, (pp. 245–256). New York: Springer.

Evaluating the External Effectiveness of the University Education in Italy

Matilde Bini

Abstract This paper aims at checking the possibility to measure the external effectiveness of course programs groups of all Italian universities, taking account of both characteristics of individuals and context factors that differently affect the Italian regions. We perform the analysis using a multilevel logistic model on data set from survey on job opportunities of the Italian graduates in 2004, conducted in 2007 by the Italian National Institute of Statistics

1 Introduction

The recent increasing unemployment rates of young graduates observed in many European countries, spurred the interest of many people to tackle the problem of the transition from university to work. An important aspect of this phenomenon is verifying if the degree of education acquired from university is adequate to the needs demanded by labor market. To do that, it is necessary to study the effectiveness of the university educational process with respect to the labor market outcomes of graduates, also named *external effectiveness*. This concept can be measured through several indicators, like the success in getting a job (Bini 1999), or a job in conformity with the education acquired by graduate (Bini 1999); or a job with adequate salary or income; or the time to get the first job (Biggeri et al. 2001).

This study deals with the most popular measure: the success in getting a job after degree. Any analysis on the effectiveness of university educational process must be performed taking account of factors that can contribute to results of the effectiveness indicators: factors pertaining individuals that are graduates' characteristics and also *internal context* factors that are universities' characteristics, like for example some indices of financial resources, number of students per course program, etc. Some studies dealing with this issue in Italy, like the surveys on job opportunities conducted by (AlamaLaurea 2009) and by the Italian National Institute of Statistics (Istat) in 2007, as well as the analysis carried out by Bini (1999), revealed that results of Northern regions are different from those of Southern regions. This means that the difference among the probabilities of getting a job (or a first job) is due not

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_3,

only to differences among individuals and among different capabilities to educate individuals universities have, but also to the economic and social differences among regions (or counties) where universities are located. As a consequence, such studies should have to be performed also including observable variables or their proxies that measure the social and economic degree of territories. Therefore, it will be possible to assert that we can make comparisons among different groups of course programs under *coeteris paribus* (i.e. same) conditions.

We remind that the external effectiveness can be defined in absolute terms (absolute effectiveness or impact analysis), that means to measure the effect of interventions with respect to a specific target; or in relative terms (relative or comparative effectiveness), that means to make comparisons among situations of many institutions. It can also be defined in two different typologies (Willms 1992; Raudenbush and Willms 1995), type A: in this case potential students are interested in comparing the results they can obtain by enrolling in different institutions, irrespective of the way such results are yielded; or alternatively type B: in this case Government is interested in assessing the "production process" in order to evaluate the ability of the institutions to exploit the available resources. Whatever definition we adopt, to analyze and make comparisons among outcomes from educational services that universities offer, it is important taking account of factors that make comparable these outcomes.

This paper aims at checking the possibility to measure the external effectiveness of course programs groups of all universities considering both the characteristics of individuals and context factors that differently affect the Italian regions, considered as *external context* factors of universities. We point out that course programs are aggregated in groups to avoid a study with a multitude of course programs having quite different denominations but same formative contents. We perform the analysis of the external effectiveness of the Italian university educational process, using data set from survey on job opportunities of the Italian graduates in 2004, conducted by Istat in 2007 (Istat 2008a). Several external context indicators (Istat 2008b) are involved in this analysis.

Because of the data set can be arranged in a hierarchical two level structure (graduates nested in groups of courses combined with universities), a multilevel logistic model is applied.

2 Data Description

The Istat data set used in the analysis is made up by two distinct stratified samples for males and females, where the strata were defined as intersections between course programs and universities. The sample size of interviews is 47,300, which is about 18% of the universe (2,60,070) of graduates. Since we want to estimate the probability to get job, the sample of 47,300 records has been reduced to 32,228 by eliminating all graduates who, at the date of the interview, did have the same

job before their degrees, and who were unemployed but at the same time were not interested in searching for a job. Data set includes two different groups of individuals: 26,570 and 20,730 graduates from degree programs existing respectively before and after past reform on teaching activity, the decree law DM n.509/99, that was effective starting from the a.y. 2001/2002. This peculiarity suggests to accomplish two separate analysis.

Our study carries out the analysis of the external effectiveness for the second sample because it refers to the recent teaching system that should affected the performance of universities in terms of capacity of educating young people to requirements of job market and also changed the propensity of people to search for a job after the degree. As regards covariates, they concern information on:

- gender ($0 =$ male; $1 =$ female);
- age ($1 = \leq 24$; $2 = 25 - 26$; $3 = 27 - 29$; $4 = \geq 30$);
- marital status ($0 =$ married; $1 =$ not married or divorced);
- address during the studies ($1 - 20 =$ number of Italian regions, 99 abroad);
- kind of the degree;
- final mark ($1 = \leq 79$; $2 = 80 - 89$; $3 = 90 - 94$; $4 = 95 - 99$; continuous if ≥ 100);
- occupational condition during the studies ($0 =$ yes; $1 =$ no);
- other studies or training after the degree ($0 =$ yes; $1 =$ no);
- graduation within institutional time ($0 =$ yes; $1 =$ no);
- military service ($0 =$ yes done; $1 =$ not yet);
- family background (0 if father doesn't have a degree; 1 if father has at least a degree);
- high school attended (9 categories).

All variables are measured at individual level, while covariates of external context added to this data set, are measured at groups of course programs combined with universities level. Each macro-economic variable characterizes one Italian region, consequently all universities belonging to the same region have the same value of the variable. These are: (a) macroeconomic measures (Gross Domestic Product per inhabitant, GDP, or productivity of labor); (b) job market measures ((youth) unemployment rate or quota of irregular labor); (c) measure of production structures (number of firms per inhabitant, average number of employees per firm); (d) innovation and technology measures (quota of innovative firms); (e) measures of the degree of culture (quota of family expenses for cultural entertainments); (f) quality of life (the poverty rate).

We assumed that graduates who continued to study two more years after their degrees (who are many in this sample), search for a job after completing their studies, so that the economics indices we use are observed for 2007 year. However, an average value of them over 3 years could also be considered as alternative. Factors pertaining internal context have not been included because they missed in this data set and it was not possible to have this information from other sources.

3 Multilevel Logistic Model

Because data have a hierarchical structure with two levels, represented by graduates (level-one units) and by groups of course programs combined with universities, labelled *class-univ* (level-two units), a multilevel regression model is used. A review of ML models in education is proposed by Grilli and Rampichini (2009). The observed response y_{ij} which measures the event of getting job (yes/no), is binomially distributed. When the response is binary, a two-level *logistic* model for the graduate i of the course program-university j, with a binomial variation at level 1 assumed, can be defined as follows (Goldstein 1995):

$$logit(\pi_{ij}) = \beta_0 + \sum_h \beta_h x_{hij} + \sum_l \delta_l z_{lj} + u_j \tag{1}$$

where π_{ij} is the probability to get job, x_{hij} is the h-th covariate measured at i-th and j-th respectively first and second levels, and z_{lj} is the l-th macro-economic covariate measured at j-th level; $u_j \sim iid\ \mathrm{N}(0,\sigma_u^2)$ is the random effect at second level. Several covariates selected from Istat questionnaire have been considered in the analysis, nevertheless only some macro-economic variables have been added in the model, like the Gross Domestic Product per inhabitant (gdp), the unemployment rate ($unempl$), the number of firms per inhabitant (n_firms), the quota of innovative firms ($innov\_firms$) and the poverty rate (q_life). They are the most representative indices of the economic and social context.

In order to have a easier interpretation of the results, some covariates have been transformed as binary by reducing the number of their categories.

4 Main Results

Two different models have been estimated firstly without and secondly with macro-economic variables, using STATA program (STATA, 2007). In Fig. 1 we show the estimates obtained from the model fitted without external context variables (Model 1). The fitting yielded that only few covariates are significant: age of graduates ($age2$), family background ($family\_backgr$) graduation within institutional time ($q1\_19$) and studies (masters or other studies) after the degree ($after\_degree$).

The unexpected negative family background variable reflects that graduates from course programs after the decree law DM n.509/99 and having parents with higher level of education, are inclined to continue their studies instead of search for a job.

The significant variance of the residual σ_u^2 confirms the presence of differences among groups of course programs of different universities. This result highlights the effects of individuals' characteristics. From the estimates it is possible to measure and compare the external effectiveness of course programs groups from different universities for any kind of graduate, besides the baseline profile with 22 years old, graduated within institutional time, who never attended master or PhD courses after degree, having a graduated parent (father). The second step of the analysis yielded

Model 1

q2_1	Coef.	std. Err.	P> | z |
age2_2	. 42844	. 0721124	0. 000
age2_3	. 9431854	. 0916032	0. 000
age2_4	. 9442713	. 1419533	0. 000
q1_19	. 4864751	. 0730985	0. 000
family_backgr	–. 338843	. 069461	0. 000
after_degree	–1. 419094	. 1174171	0. 000
_cons	–1. 523007	. 2194461	0. 000
sigma_u	. 6899792	. 0895518	

Fig. 1 Estimates of model without external context variables

Model 2

q2_1	coef.	std. Err.	P> | z |
Level 1			
age2_2	. 42086	. 0719434	0. 000
age2_3	. 9428713	. 0913991	0. 000
age2_4	. 9471336	. 1416367	0. 000
family_backgr	. 3428294	. 0694086	0. 000
q1_19	. 4924442	. 0730578	0. 000
after_degree	–1. 420495	. 1173548	0. 000
_cons	. 9413027	. 7402302	0. 204
Level 2			
gdp_07	. 4497361	. 1939072	0. 020
unempl_07	–1. 3617416	. 1260271	0. 004
sigma_u	. 4232194	. 0638389	

Model 3

q2_1	coef.	std. Err.	P> | z |
Level 1			
age2_2	. 4240396	. 0719473	0. 000
age2_3	. 9451951	. 0914018	0. 000
age2_4	. 9441328	. 1415598	0. 000
family_backgr	–. 3404437	. 0694056	0. 000
q1_19	. 4921052	. 0730486	0. 000
after_degree	–1. 419544	. 1173609	0. 000
_cons	2. 182026	. 3099628	0. 000
Level 2			
n_firms_07	. 0292658	. 0132538	0. 027
unempl_07	–. 4859348	. 0978811	0. 000
sigma_u	. 4367879	. 0634571	

Fig. 2 Estimates of models with external context variables

significant estimates for Gross Domestic Product per inhabitant (gdp_07), unemployment rate ($unempl\_07$) and poverty rate (q_life_07); while variables such as the number of firms per inhabitant (n_firms_07) and the quota of innovative firms ($innov\_firms\_07$) revealed to be significant provided if they are included as alternative to the GDP. To give an example, Fig. 2 shows the estimates of two models, Models 2 and 3, which perform the probability to get job taking account of the GDP per inhabitant or the number of firms per inhabitant, and the unemployment rate. They are the most important indices representing the degree of economic and job market development of a territory. Hence, it will be possible to measure and compare the external effectiveness of course programs groups from different universities for any kind of graduate, but under more comparable conditions due to the presence of effects of economic and market conditions where universities are located.

4.1 Context Variables Effects

To highlight the contribution of external variables, it is possible to construct a very simple index that measures the difference (in percentage) of the second level variability between the two different models, respectively free of external variables and under coeteris paribus:

$$I_{\Delta} = \left(\frac{\sigma^2_{u,without} - \sigma^2_{u,with}}{\sigma^2_{u,without}} \right) * 100 \tag{2}$$

In other words, this index represents the percentage of variability "explained" by context factors. Here in this analysis, since the second level variances for example of model 1 and model 2 are respectively 0.68 and 0.42, the value of index is equal to 38.2%. This means that the 38.2% of the differences among groups of course programs is due to the presence of GDP and unemployment rate in the model. Moreover, the importance of external variables in the analysis is also evaluated if we compare the estimation of the probability to get job for the baseline graduate, according to a classification of five different typologies of universities. Let consider this probability to be:

$$\pi_{ij} = \frac{exp(\beta_0 + \sum_h \beta_h x_{hij} + \sum_l \delta_l z_{lj} + u_j)}{1 + exp(\beta_0 + \sum_h \beta_h x_{hij} + \sum_l \delta_l z_{lj} + u_j)} \tag{3}$$

Under the standard assumption of normality of the distribution of the residual u_j, we define the following classification of the universities using the estimated variance σ^2_u: (1) very good universities if $u_j = 2\sigma_u$; (2) good universities if $u_j = \sigma_u$; (3) medium level universities if $u_j = 0$; (4) bad universities if $u_j = -\sigma_u$; (5) very bad universities if $u_j = -2\sigma_u$. We substitute these five values of u_j in the linear predictor $\beta_0 + \sum_h \beta_h x_{hij} + \sum_l \delta_l z_{lj} + u_j$.

In Fig. 3 we plot the estimated probabilities of getting job for all five categories of universities with respect to all the regional GDPs values observed for the year 2007. The estimates are obtained when the unemployment rate is fixed as same value (i.e. average rate value) for all regions. Universities located in the same region have the same value of GDP, as well as several regions can have quite similar values of GDP. From this Figure it is possible to notice that the trajectories are separate and parallel across all regions. This means that, for example, a very good university is the same independently from where it is located; belonging to a category can not be affected by the localization. But in terms of comparisons among units, the localization affects the external effectiveness of universities: for example, very bad universities located in a region having the highest GDP value (3.3), are similar to good universities with lower level of GDP (1.6). Moreover, divergences among categories decreases as we move towards regions with higher level of economic development. Finally, also when we focus on just one group of course programs, namely Economics, we obtain same conclusions.

In Fig. 4, probabilities to get job for graduates in Economics, according the mentioned classification are plotted as regards four Italian regions Lombardia, Tuscany, Campania and Sicily, given as an example. Here results show that bad groups of course programs located in the best regions defined in terms of economic and social development, are always much better than best universities located in bad regions. This means that comparisons must be done also including factors characterizing the regions and affecting external outcomes of the universities.

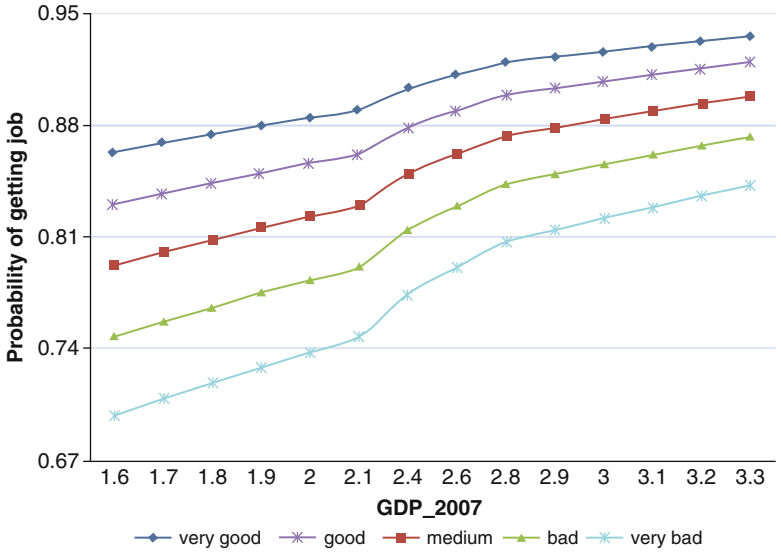


Fig. 3 Probabilities of getting job: a classification

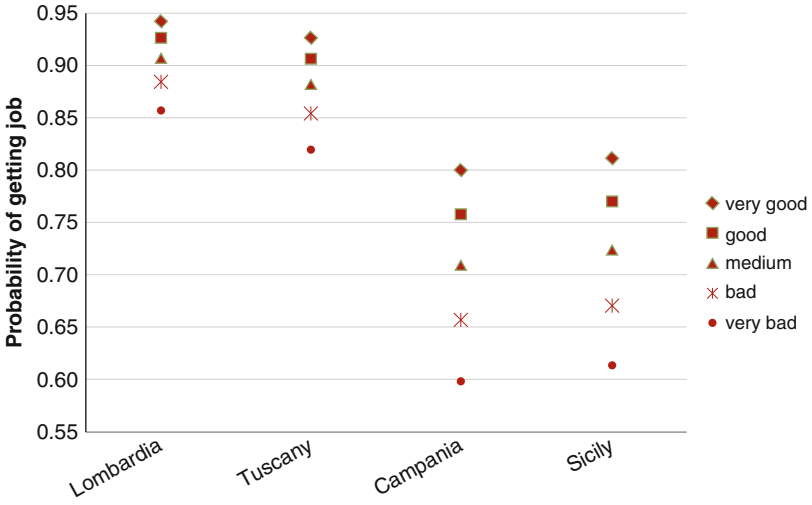


Fig. 4 Probabilities of getting job for graduates in Economics: a classification

5 Conclusive Remarks

This work analysis the external effectiveness considering all these following aspects: (1) the study of the influence of the individuals' factors that affect their probabilities to get job; (2) the evaluation whether and how much the differences among economics and social territories are important in the probability to get a job and

influence the choices of people in searching for a job; this means to evaluate factors of the external context; (3) the evaluation of the differences among course programs of universities with respect to the probability to get a job, taking account of the fact that the relationships among variables at individual level can vary according to the course programs of universities. Results pertaining this aspect give more properly a measure of the relative effectiveness of universities. This could help students to assess which university or which course program among universities offers the best results in terms of performance and job opportunities. The use of context characteristics improve results in the analysis of ranking of universities (or course programs) but there is a need to improve this measure including internal contexts (characteristics of institutions). We also remark that context variables are observed only at region level while there is a need to have information about territories at a more disaggregate level (i.e. counties).

References

AlamaLaurea. (2009, Aprile 1). *XI Rapporto sulla condizione occupazionale dei laureati. Occupazione e occupabilità dei laureati a dieci anni dalla Dichiarazione di Bologna. A. Cammelli (a cura di)*. Bologna: AlmaLaurea.

Biggeri, L., Bini, M., & Grilli, L. (2001). The transition from university to work: A multilevel approach to the analysis of the time to obtain the first job. *Journal of the Royal Statistical Society series A*, *164*(part2), 293–305.

Bini, M. (March 1999). *Valutazione della efficacia dell'Istruzione universitaria rispetto al mercato del lavoro*. Research Report n.3/99 (pp. 1–100). Roma: Osservatorio per la Valutazione del Sistema Universitario-MIUR.

Goldstein, H. (1995). *Inserimento professionale dei laureati. Indagine 2007*. London: Edward Arnold.

Grilli, L., & Rampichini, C. (2009). Multilevel models for the evaluation of educational institutions: A review. In M. Bini, P. Monari, D. Piccolo, & L. Salmaso (Eds.), *Statistical methods and models for the evaluation of educational services and product's quality*. Phisica-Verlag, Berlin Heidelberg, 61–79.

ISTAT. (2008a). *Inserimento professionale dei laureati. Indagine 2007*. Roma: ISTAT.

ISTAT. (2008b). *100 Statistiche per il Paese. Indicatori per conoscere e valutare. A G. Barbieri, S. Cruciani, A. Ferrara (a cura di)*. Stampa CSR - Maggio.

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, *20*, 307–335.

STATA Release 10.0. (2007). *Data analysis and statistical software for professionals*. URL: http://www.stata.com/.

Willms, J. D. (1992). *Monitoring school performance: A guide educators*. London: Falmer.

Analyzing Research Potential through Redundancy Analysis: the case of the Italian University System

Cristina Davino, Francesco Palumbo, and Domenico Vistocco

Abstract The paper proposes a multivariate approach to study the dependence of the scientific productivity on the human research potential in the Italian University system. In spite of the heterogeneity of the system, Redundancy Analysis is exploited to analyse the University research system as a whole. The proposed approach is embedded in an exploratory data analysis framework.

1 Introduction

Nowadays, the evaluation of University systems plays a crucial role in the quality certification and it is a key element to be present and competitive all over the world. Because of the increasing gap between reduced public resources and research cost raising, the assessment of scientific research communities has become even more important. As a consequence, research assessment systems concern: institutional evaluation aiming to get researchers' skills and capabilities valued, certification of research potentiality and financial evaluation, to allocate founds according to.

A scientific research community can be evaluated through objective or subjective indicators. The former are derived from the quantification and evaluation of the research products (see for example citation data such as the Impact Factor). The latter are based on judgments by scientists, named peer reviewers. Both approaches are directed to provide a ranking of the considered research communities through suitable and different synthesis methods.

In the European countries, the Research Assessment Exercise (RAE 2008) is one of the most consolidated and formalized assessment process for the evaluation of the quality of research products. In Italy, even though several evaluation processes have been recommended, a unique and formalized approach has not been regulated, yet. The Italian Department of Education (MUR), the National Committee for University System Evaluation (CNVSU) and the Italian Committee for Research Evaluation (CIVR) represent the most important authorities in the field.

The approach proposed in this paper is inspired to the model for the allocation of funding to Universities (FFO) proposed by the CNVSU (2005). According to this

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_4,

model, funding are allocated to Universities on the basis of the ranking deriving from the *research potential*, an indicator which expresses the theoretical research capability of the Universities. Such indicator synthesizes both the total number of researchers belonging to the University and their ability in promoting and obtaining funding. The *research success ability* has been defined by CNVSU on the basis of the total number of researchers involved in PRIN projects (Programmi di Ricerca di Interesse Nazionale) that received a positive score (not necessarily funded).

Starting from the CNVSU model, this paper presents a multivariate approach able to provide a global analysis of the Italian research potential and to take into account the relationships among the scientific areas. Among the several statistical multivariate approaches available to this aim, the paper exploits the Redundancy Analysis (Wollenberg 1977), that can be viewed as a variant of Canonical Correlation Analysis in case of asymmetric or dependence structures.

The proposed approach permits to highlight the risk related to the construction of a ranking of the Italian Universities obtained without considering the peculiarity of Universities and of the scientific areas.

Finally, a sensitivity analysis is carried out to show and evaluate how much the dependence between the obtained and expected scientific results is affected by the selected Universities and/or scientific sectors. The analysis is conducted on the 2007 MUR data available at http://prin.cineca.it.

2 The PRIN Dataset

The evaluation of the effectiveness of research projects is based on the analysis of the data related to the applications for national research projects (PRIN) funding.

Every year, Italian University researchers can apply for the national research funding PRIN. As a common practice, projects are jointly submitted by two or more groups from the same or different Institutions. Although different scientific sectors can be involved in a project, a leader sector have to be indicated in the application form. Each scientific area has its own board, a national committee supervises the activity of the different boards. The final outcome of the evaluation process consists of the following three conditions: granted, positive but not granted, rejected.

This paper aims at comparing the number of researchers participating in projects that got a positive evaluation (granted or not) with respect to the available human resources, taking into account the 14 scientific areas. These are:

Area 01: Mathematics and Information Science	Area 02: Physics
Area 03: Chemistry	Area 04: Geology
Area 05: Biology	Area 06: Medicine
Area 07: Agriculture	Area 08: Civil Engineering and Architecture

Area 09: Industrial Engineering	Area 10: Literature and Art History
Area 11: History, Philosophy and Pedagogy	Area 12: Law
Area 13: Economics and Statistics	Area 14: Social and Political Science

In this specific context, let us consider two *Universities*×*scientific areas* data tables named: *Research Potential* and *Success Ability*, respectively. As generic element, the former contains the counts of the full, associate and assistant professors for a given University into one scientific area (POT); the latter has the same structure and contains the counts of researchers that have participated to a PRIN project, which gained a positive evaluation (POSTOT). In the following we assume that the two data tables represent quantitative variables and that the counts respectively represent potentiality and success indicators with respect to each scientific area and measured for any University. This assumption does not affect the whole analysis approach and permits to resort to statistical methods for continuous variables.

In order to describe the research potential and the success ability distributions with respect to both scientific sectors and Universities, univariate and bivariate simple statistics can be carried out. Figure 1 reveals a high heterogeneity level among sectors, both for research potential and success ability. This heterogeneity depends on the different peculiarities of scientific sectors as a whole but also from the different size and the different scientific features of the Universities.

It is straightforward to expect the presence of strong correlations among research potential (POT) and success ability (POSTOT), as confirmed by the following values for the different scientific sectors:

Area 01: 0.945; Area 02: 0.968; Area 03: 0.977; Area 04: 0.942; Area 05: 0.968; Area 06: 0.954; Area 07: 0.982; Area 08: 0.980; Area 09: 0.988; Area 10: 0.901; Area 11: 0.946; Area 12: 0.928; Area 13: 0.894; Area 14: 0.939.

Under these conditions, any global analysis, such as the ranking of the Universities according to their scientific productivity becomes a ticklish problem.

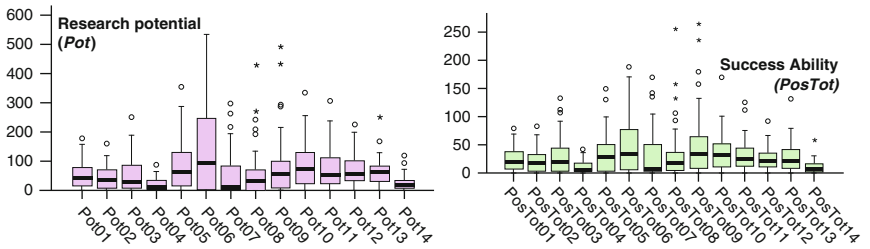


Fig. 1 Boxplot of Universities distribution in the scientific sectors for research potential (*left*) and success ability (*right*)

3 The Redundancy Analysis Approach

3.1 The Method

Redundancy Analysis (RA) Wollenberg (1977) studies the dependence of one battery of variables $Y = \{Y_1, \ldots, Y_p\}$ on another battery $X = \{X_1, \ldots, X_p\}$. Under the condition that assumes the Italian University system as a whole, success ability (POSTOT) represents the dependent variable set Y and research potential (POT) is the explicative one X.

RA can be considered as a reduced rank regression or a principal component analysis of the projections of the Y on the space spanned by the X. It faces the previously highlighted issues in the interset correlation analysis: different size of the Universities, different scientific peculiarities of the Universities and relationships among the areas. Scaling variables to unit variance allows us to overcome the heterogeneity within scientific areas.

In the algebraic notation, $\mathbf{X}$ and $\mathbf{Y}$ respectively represent the variable sets X and Y, they have the same order $N \times p$ and are centered and scaled to unit variance where $N = 58$ indicates the number of Italian Universities involved into the analysis, and $p = 14$ represents the number of scientific areas. RA aims at calculating the variates $\xi_r = \mathbf{X}\mathbf{w}_r$, for $r = 1, 2, \ldots, p$, and with unit variance such that the sum of squared correlations of the Y variables with the variates is maximal.

The correlations of the Y with the variate ξ are given by the column vector $(1/N)\mathbf{Y}'\mathbf{X}\mathbf{W}$; the sum of squared correlations is equal to the minor product moment.

Formally the problem consists in maximizing the quantity:

$$Q = \frac{1}{N^2}\mathbf{W}'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{W}, \tag{1}$$

under the constraint $\frac{1}{N}\mathbf{W}'\mathbf{X}'\mathbf{X}\mathbf{W} = \mathbf{I}$.

3.2 The Results

The present paper studies the proportion of the total variance of the success ability by a canonical multivariate analysis of the research potential.

The decomposition of the percentage of inertia of POSTOT resulting from RA attributes 94% of such variability to the relation between POSTOT and POT (constrained analysis) and it confirms that much of the variability of Success Ability is explained by the number of researchers. Nevertheless, the methods allows to quantify the intensity of such a relation. The remaining 6% can be considered as the residual part of variability, after removing redundancy due to the effect of POT on POSTOT. Notwithstanding the reduced unconstrained part, it can be interesting to highlight if some scientific area or some Universities is characterized by a particular ability, independently from the number of researchers belonging to it.

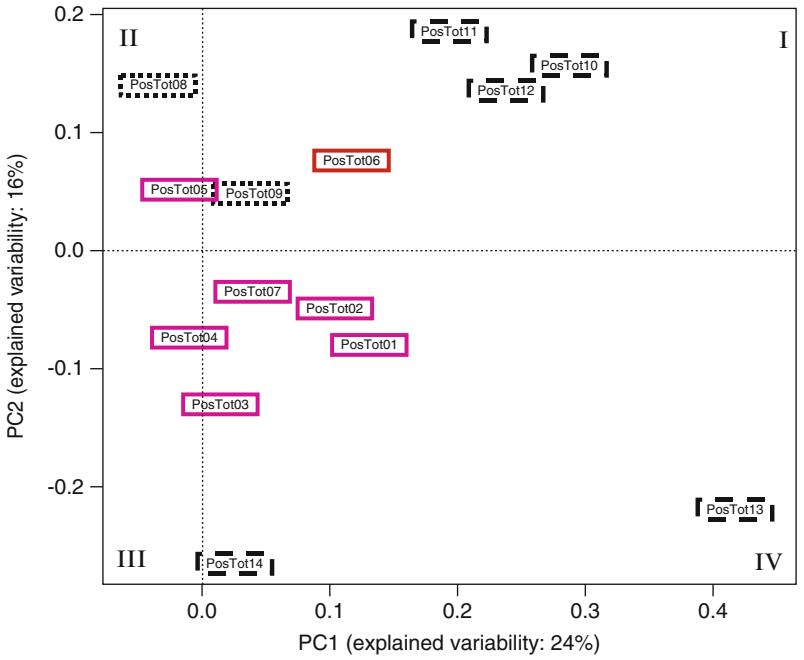


Fig. 2 First factorial plan: space of the variables

Figure 2 represents the variables POSTOT on the first factorial plan after removing redundancy. The interpretation of the plan follows the same rules of a classical Principal Component Analysis.

The first remark deriving from the analysis is that it is not affected from relationships among the related areas and it is able to show the heterogeneity among the areas. The plane separates related areas highlighted with the same line shape, particularly for Human sciences (dashed line) and to engineering (dotted line). Hard sciences still remain very close. This can be partially explained from the strong cooperation characterizing such sciences.

It is worth to notice how much area 13 places far from the other sectors, showing a high capability to excel.

The first factorial plan explains 39.6% of the total variability. In order to reach a reasonable amount of explained variance, the second factorial plan (spanned by axes 3 and 4) is considered, too (Fig. 3). Such two plans give 61.2% of the total variability. The low level of explained variance strengthen the belief that the method is able to catch the actual capability and peculiarity of each University in being positively evaluated with respect to the 14 scientific areas. The second factorial plan (Fig. 3) reveals a capability to excel by area 12 and area 4.

The configuration of the statistical units with respect to the first two factors (Fig. 4) shows that Universities are not discriminated with respect to the size.

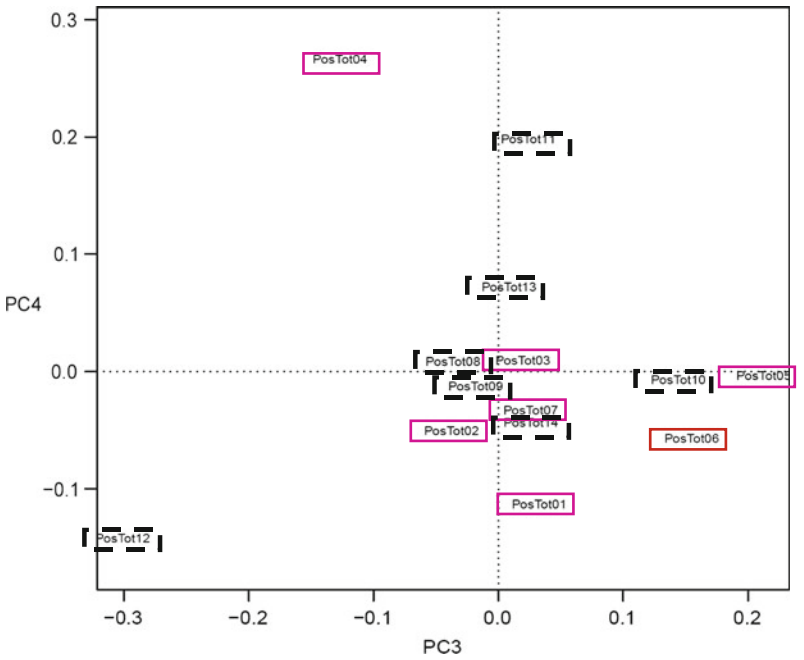


Fig. 3 Second factorial plan: space of the variables

Anyway, most of the biggest Universities (NAP1, BA, Roma, PR, ...), lies far from the origin, they are characterized by a capability to excel in one or more areas.

A possible ordering of the Universities is showed in Fig. 5: Universities are ranked according to their distances from the origin on the first 4 factors. Some remarks can be drawn:

- A short distance from the origin can be interpreted as low capability to excel in whatever area while a high distance reveals that the University is characterized by a particular ability or inability. It is worth of notice that the aim of the paper is not to provide a ranking of Universities from those with a high ability to excel to the ones with the worst research ability but to highlight peculiarities in the research aptitude beyond the research potential.
- The dimension of each University does not affect the analysis (biggest Universities are represented in rectangles).
- Specialized Universities such as polytechnics (PoliMi, PoliBa, PoliTo) do not separate from the others and they do not excel with respect to the other Universities.

In order to reinforce the above results and remarks and to verify the stability of the detected patterns structures, a sensitivity analysis is advisable.

It is well-known that Sensitivity Analysis Saltelli et al. (2008) aims to identify the contribution of each factor involved in the construction of a composite indicator

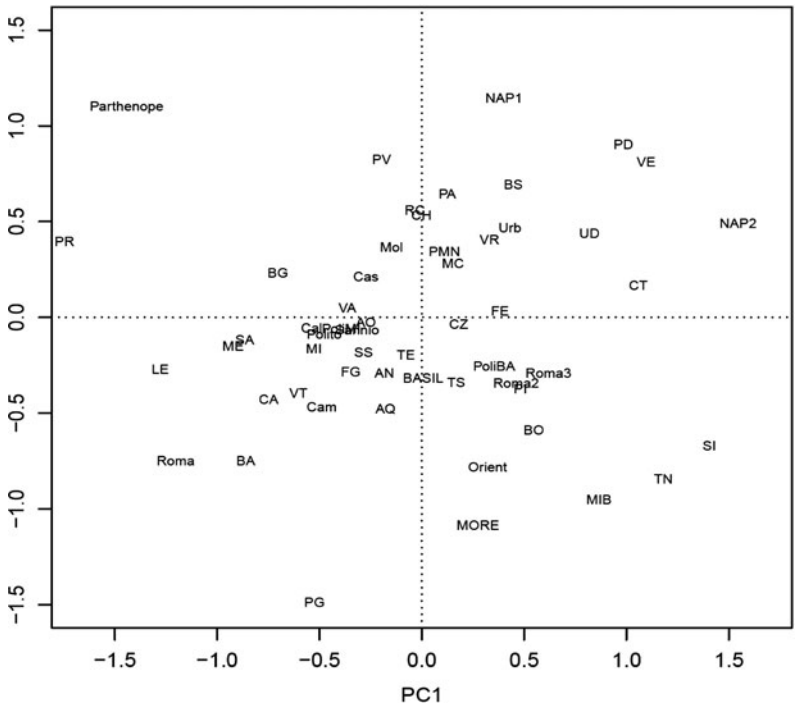


Fig. 4 First factorial plane of RA: space of the of the units

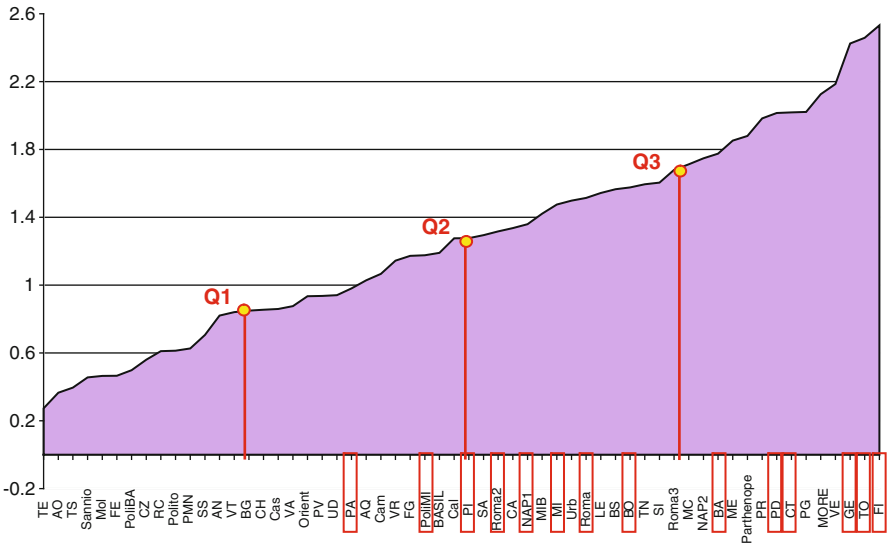


Fig. 5 Distance (*y-axis*) of each University (*x-axis*) computed on the first 4 factors

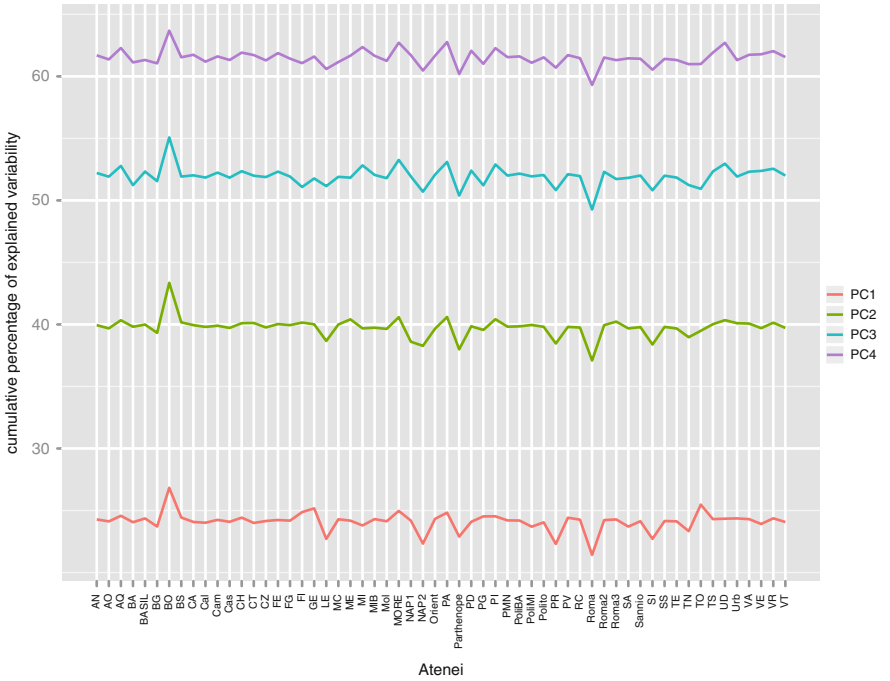


Fig. 6 Stability of the dependence structure

(weighting schemes, aggregation methods, etc.) on the uncertainty of the composite indicator. In the present paper, the Sensitivity Analysis concept is borrowed and adopted to the RA results. A first step is performed using a leaving one-out procedure, namely carrying out several RA counting out a unit at a time. In Fig. 6 the cumulative percentage of variability explained on the first four factors is shown on the vertical axis with respect to the excluded University shown on the horizontal axis. Peaks in the distribution of the explained variability reveal that the corresponding excluded unit can cause troubles to the stability of the dependence structure. The stability of results may be also explored by looking at the correlations between each pair of factors. Moreover, further analysis could include an evaluation of the role played by the scientific sectors. Following the approach introduced to evaluate the stability of the dependence structure from the units, a leaving one-out procedure can be performed carrying out several RA counting out a scientific sector at a time.

References

Roberts, G. (2003). Review of research assessment: Report by Sir Gareth Roberts to the UK funding bodies issued for consultation May 2003, Ref 2003/22, London Higher education Funding Councils.

CNVSU – Comitato nazionale per la valutazione del sistema universitario (2005), *Il modello per la ripartizione del fondo di finanziamento ordinario (FFO) all'interno del sistema universitario: riflessioni a valle dell'applicazione sperimentale prevista dal D.M. 28 luglio 2004*, Doc. n. 4/05, Ministero dell'Istruzione, dell'Universitá e della Ricerca, Roma.

RAE 2008. (2008). *Research assessment exercise: The outcome*. Retrieved December, 2008, Ref RAE 01/2008, from http://submissions.rae.ac.uk/results/outstore/RAEOutcomeFull.pdf.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gabelli, D., Saisana, M., & Tarantola, S. (2008). *Global sensitivity analysis. The Primer*. John Wiley & sons, England.

Saito, T., & Yadohisa, H. (2005). *Data analysis of asymmetric structures*. New York: Marcel Dekker.

van den Wollenberg, A. L., (1977). Redundancy analysis: An alternative for canonical correlation analysis. *Psychometrika*, *42*(2), 207–219.

A Participative Process for the Definition of a Human Capital Indicator*

Luigi Fabbris, Giovanna Boccuzzo, Maria Cristiana Martini, and Manuela Scioni

Abstract In this paper, we discuss a method for defining the hierarchical structure of a composite indicator of graduate human capital that could be used to measure the educational effectiveness of Italian universities. The structure and weights of the dimensions of graduate human capital, and the set and weights of the elementary indicators, were determined using a three-round Delphi-like procedure. We contacted the rectors, the presidents of the evaluation boards and other qualified professors at Italian universities, as well as representatives of worker unions and entrepreneur associations. Our exercise shows that most dimensions of graduate human capital are related to the educational role of universities and that weights and indicators of the dimensions can plausibly be measured with the participation of the concerned individuals.

1 A Human Capital Indicator

The concept of effectiveness encompasses relationships between the results of service delivery and the aims that the service itself was expected to attain. The application of effectiveness principles requires that service objectives are stated in advance in operational terms and that effectiveness indicators are specified for measurement purposes. A service may, therefore, be fully, partially, or not at all effective according to the level of purpose attainment. These principles apply either to the whole system of educational services or to its component parts, i.e. universities, faculties, study programmes and even single courses.

The concept of educational effectiveness echoes students' learning, so an educational structure can be said to be effective if the students who were subject to that structure learned as expected (Scheerens and Bosker 1997). In common with

*This research was supported by a grant from the Italian Ministry of Education, University and Research (PRIN 2007) as well as by a grant from the University of Padua, Italy (Athenaeum Project 2008).

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_5,

Hanushek (1997), we distinguish between internal and external effectiveness of university education. If assessment is conducted just within the university, we are confined to internal effectiveness; if it is related to civil society, and in particular the workplace, the concept is broadened to external effectiveness. The problem with effectiveness is that there is no general agreement on how it should be measured.

In 2009, measurement of the effectiveness of universities in Italy became a critical issue when a law decree stated that a quota of annual funds from the Ministry should be allotted to universities, *inter alia*, according to their educational effectiveness.

We present a method for the construction of a hierarchical composite indicator of graduate human capital. Using a participative process, we first define the main dimensions of the construct. We then employ a set of indicators for each dimension. We argue that such an indicator can be assumed to be a measure of the external effectiveness of tertiary education. We present and discuss our definitional model in Sect. 2 and present our results of the participative survey process in Sect. 3.

2 A Participative Method to Define Human Capital

Human capital (HC) is an evolving concept. A definition put forward by the OECD (1998) based on Becker's and other economists' ideas (Becker 1993) describes HC as comprising the knowledge, skills, competencies and other attributes embodied in individuals that are relevant to economic activity. We, however, consider that a definition that relates a person's competence just to economic values is incomplete and suggest that the definition should be broadened to include societal issues (see also Coleman 1990). So, in terms of graduate human capital (GHC), we propose that it be defined as an individual's capacity to add value to society as a worker and as a citizen.

If we consider graduates educated in different university structures and measure their learning outcomes, we can say that the learning mean values reflect the structures' differential effects on the graduates, all other factors being equal. The concept of a person's HC is, therefore, extended to encompass the structures that generate it. In this way, educational structures can be ranked or scored according to their effectiveness. Thus, measures of graduates' outcomes can be assumed to be mirror values of GHC provided by the attended educational structure.

We propose a hierarchical procedure for defining a GHC indicator. The procedure was experimentally tested using Italian tertiary education stakeholders: rectors, presidents of the evaluation boards and other qualified professors of Italian universities, as well as representatives of worker unions and entrepreneur associations. These experts provide diverse views of GHC, according to their membership to university, which contributes to the HC formation, to labour market or to society, where graduates are expected to spend their HC.

A Delphi-Shang approach (Ford 1975; Fabbris et al. 2008) was adopted; this approach involves a dialogic surveying process between the researcher and the expert group and a progressive focus on the topical issue. At any Delphi iteration,

feedback on the responses obtained was returned to the experts, and a revised, more sharpened set of questions was then asked.

In the first step, experts were asked to indicate what they considered the main dimensions of a GHC indicator to measure educational effectiveness. The answers, provided in an open-ended format, were then assessed, discussed and systematized by the research group, leading to the eight common dimensions reported in Sect. 3. In the second step, experts were requested to rate the relative importance of the eight dimensions, and to propose a set of possible indicators for each dimension. These indicators were then discussed and reduced to 60 indicators, either 8 or 6 for each dimension (see Table 3). The final set of indicators was selected on the basis of factuality of measures, exhaustiveness and non-redundancy of contents. The final step focused on adjusting and confirming the weights previously assigned to the eight dimensions and rating the importance of the indicators.

Experts were contacted by email and invited to complete a web questionnaire. The survey lasted from mid-January 2009 to the beginning of March 2009. Sixty two individuals, of whom 51 were academics, completed the first questionnaire. Thirty one experts completed the second questionnaire, with 22 being academics. The third questionnaire was completed by 30 experts, 23 of whom were academics.

To elicit dimensions' weights, experts were randomly assigned to two data collection systems: the "budget system", that is, the distribution of a 100 points budget over the alternatives, and the ranking of dimensions according to their relevance for GHC. In the first system, weights were estimated as the arithmetic mean of the budgets assigned by the respondents to every dimension. In the second case, a multivariate analysis of preferences was performed. The preference estimates of dimensions were ordered in a dominance matrix $\mathbf{P} = \{p_{ij}\ (i, j = 1, \ldots, k)\}$, where k is the number of assessed dimensions and p_{ij} is the relative frequency with which dimension i dominates j $(i \neq j = 1, \ldots, k)$, which may also be written as $i \succ j$. In this matrix, $p_{ii} = 0$ and $p_{ji} = 1 - p_{ij}$ (Fabbris 2010). It is square and irreducible.

Weights were estimated with the eigenvector associated to the first positive eigenvalue of matrix $\mathbf{P}$ (Saaty 1983):

$$\mathbf{Pv} = \lambda_{\max}\mathbf{v} \tag{1}$$

subject to $\mathbf{v}'\mathbf{v} = 1$. Weights w_i of dimension $i\,(i = 1, \ldots, k)$ are proportional to v_i such that $\sum_{i=1}^{k} w_i = 1$, that is $w_i = v_i / \sum v_i$.

Weights obtained with the two methods were then synthesized using the following formula:

$$\hat{w}_i = \alpha_i\ w_{i1} + (1 - \alpha_i)\ w_{i2} \tag{2}$$

where:

- w_{i1} is the estimate of the i-th component weight obtained with the 100-point budget method.
- w_{i2} is the estimate of the i-th component weight obtained with the eigenvector method.

- $\hat{w}_i$ is a combined estimation of the i-th component's weight obtained averaging the two elicitation methods.
- $\alpha_i \propto \text{var}(w_{i1})^{-1}$, then $\alpha_i = \text{var}(w_{i2})/[\text{var}(w_{i1}) + \text{var}(w_{i2})]$ so that the more stable weight is preferred.

3 Weight Estimation of Dimensions and Indicators

The eight dimensions of GHC as derived from the systematization of the experts' answers were as follows:

1. *Culture*: Ability to correctly use terms and notions of the science and humanities culture possessed by the majority of highly educated people.
2. *Basic technical competencies*: Ability to use foreign languages and computer tools needed to obtain a job consistent with the possession of a university degree.
3. *Technical-specific competencies, problem solving skills*: Ability to solve a wide range of work problems by drawing upon specific professional skills.
4. *Learning and knowledge transfer skills*: Ability to understand the meaning of professional experiences, organize these experiences and disseminate them to a wider audience.
5. *Relational and communication skills*: Ability to generate, maintain and promote social relationships and to effectively communicate with others.
6. *Organizational and entrepreneurial skills*: Ability to manage work groups and to carry out complex activities, possibly by innovating strategic targets and management standards.
7. *Work-oriented personality*: Personal skills such as empathy with work colleagues.
8. *Social ethics and value endorsement*: Professional behaviours and attitudes arising from an ethical code, which favour collective, rather than individual, growth.

The first three dimensions – culture, basic and technical-specific competencies – are more directly related to the graduate's study programme. The ability to learn and transfer knowledge and relational, communication, organizational and entrepreneurial skills are defined as "soft-skills" or cross-occupation competencies because they can be used to deal with almost any work problem. Possessing a job-oriented personality, however, as well as social ethics, are personal characteristics relevant not only for the professional success of the graduate but also for his/her social well-being.

Tables 1 and 2 present the estimates of the dimensions' weights and the respective standard errors (s.e.), both for the whole sample and the two categories of academic and labour experts. The weights are 8.9% for the basic competencies of computer and foreign language skills, 10.9% for culture and 17.5% for technical-specific competencies related to professional problem solving. Culture, basic and technical-specific competencies – knowledge and skills that universities are expected to form

Table 1 Estimates of percentage weights for eight dimensions of human capital of graduates according to data collection system

Dimensions	Budget (n = 19)	s.e.	Order (n = 21)	s.e.	Total (n = 40[a])
Culture	12.6	7.6	9.5	6.4	10.9
Basic competencies	11.3	7.3	7.3	5.7	8.9
Technical-specific, problem solving skills	16.4	8.5	18.3	8.4	17.5
Learning and knowledge transfer skills	11.6	7.3	15.7	7.9	13.7
Relational and communication skills	10.5	7.0	11.4	6.9	11.1
Organizational and entrepreneurial skills	12.9	7.7	14.1	7.6	13.6
Work-oriented personality	12.1	7.5	10.5	6.7	11.3
Social ethics and value endorsement	12.6	7.6	13.1	7.4	13.0
Total	100.0		100.0		100.0

[a]The total amount of responses (n = 40) is due to 30 experts who answered to the third phase questionnaire and 10 other experts who answered to the second questionnaire but did not participate to the third phase.

Table 2 Estimates of percentage weights for eight dimensions of human capital of graduates according to expert membership

Dimensions	Total (n = 40)	Academia (n = 31)	s.e.	Labour (n = 9)	s.e.
Culture	10.9	11.7	5.8	5.1	7.3
Basic competencies	8.9	9.5	5.3	5.3	7.5
Technical-specific, problem solving skills	17.5	16.9	6.7	20.9	13.5
Learning and knowledge transfer skills	13.7	13.3	6.1	15.9	12.2
Relational and communication skills	11.1	11.0	5.6	12.3	10.9
Organizational and entrepreneurial skills	13.6	13.4	6.1	15.2	12.0
Work-oriented personality	11.3	11.5	5.7	10.2	10.1
Social ethics and value endorsement	13.0	12.6	6.0	15.1	11.9
Total	100.0	100.0		100.0	

in a structural way – ideally account for about 37% of a graduate's potential for work.

Soft skills, such as the aptitude for life-long learning, the ability to transfer learned experience in dealing with people and using technical knowledge in work and social settings account for 38% of a graduate's potential for work. Personal traits relevant to the workplace and moral attitudes constitute another 24% of GHC.

Opinions expressed by both the academics and the labour experts seldom diverged. Academic opinion prevailed in the general ranking, due to the larger size of the group. Both those from academia and labour groups considered technical-specific competencies as the most important employment attributes of graduates, whereas they assigned different importance to culture (Table 2). Presumably, university teachers envisage prestigious roles for graduates and assign greater value to culture (11.7%), whereas labour experts, who scored this component lower (5.1%) consider culture less important in relation to the technical positions that

new graduates take up. We assume that entrepreneurs would assign culture a higher score for more highly qualified positions than those expected from new graduates.

Basic competencies were assigned low weights especially by labour experts, possibly due to computer and language skills being considered automatic even for entering the university. The t-test did not reveal significant differences between the two groups in terms of the importance assigned to soft skills and personal attitudes.

Table 3 presents more detailed results derived from the analysis of the weights attributed to each indicator of the eight dimensions. In relation to culture, both

Table 3 Percentage weight estimates of the most important indicators for each dimension of human capital of graduates (just first half of the proposed items is reported)

Components	Total (n = 40)	Academia (n = 31)	Labour (n = 9)
Culture (eight items)			
Reading at least one book not pertaining to own job in the last year	17.2	16.7	19.1
Reading a national newspaper everyday, or almost everyday	15.6	16.5	13.5
Performing extracurricular musical, artistic and cultural activities	15.6	16.5	12.8
"Liceo" type of high school	14.7	14.1	16.5
Basic competencies (eight items)			
Ability to read and converse by phone in English	18.6	19.3	18.1
Basic computing skills (text writing, e-mail, spreadsheets, database)	17.1	17.1	15.7
Knowledge of at least two foreign languages	14.6	13.3	20.1
Ability to use internet to retrieve job-related information	13.6	13.1	15.7
Technical-specific competencies, problem solving skills (eight items)			
Frequent use of technical-specific competencies at work	19.1	17.8	25.4
Frequent use of disciplinary *forma mentis* at work	16.0	16.0	16.4
Time elapsed between recruitment and first promotion	12.4	14.6	4.5
Past or present professional counseling activities	12.3	11.6	14.2
Learning and knowledge-transfer skills (eight items)			
Attitude compatible with motivating and helping colleagues	22.5	21.3	30.8
Willingness to deal with new problems and to use new tools at work	20.6	19.7	25.2
Ability to teach fellow workers and to support young colleagues	19.7	18.6	26.2
Ability to summarize professional contents (texts, conferences, etc.)	14.4	15.9	6.1
Relational and communication skills (eight items)			
Inclination to act as an intermediary between opposing interests, to negotiate	17.9	16.7	23.5
Ability to put forward arguments for personal ideas in public	16.9	16.7	17.5
Ability to deal with both customers and suppliers	16.0	16.0	15.9
Ability to effectively write research reports or activities projects	15.7	16.1	14.0

(Continued)

Table 3 (Continued)

Components	Total (n = 40)	Academia (n = 31)	Labour (n = 9)
Organizational and entrepreneurial skills (eight items)			
Experience in conceiving or carrying out a project	19.9	19.7	22.0
Having held a position of responsibility at work (e.g. team manager)	18.1	17.4	18.4
Ability to motivate colleagues at work	16.6	13.8	36.1
Having set up in self-employment (alone or in partnership)	15.9	17.9	4.2
Work-oriented personality (six items)			
Personal motivation, draws stimulation from results of work	25.5	23.7	33.0
Availability to work after hours or during the week-end	17.7	18.4	15.0
Precision, accuracy and diligence	17.5	17.4	17.4
Social ethics and value endorsement (six items)			
Propensity to give credit to other people at work	24.0	22.2	32.5
Compliance with internal rules and the law	21.6	21.6	21.8
Assigns importance to social equity	20.5	21.2	18.8

academics and labour experts placed most value on graduates having read at least one book in the past year not pertaining to their own speciality. Labour experts assigned this indicator a weight equal to 19.1%, whereas academics assigned it a value of 16.7%.

Labour experts proposed the following indicators as adequate measures of basic competencies: knowledge of at least two foreign languages (20.1%), ability to read and converse by telephone in English (18.1%), ability to use the internet and retrieve information for a particular job (15.7%), and a reasonably good level of knowledge of computers (15.7%). Overall, these four indicators account for approximately 70% of the preferences. The importance that the labour experts place on the possession of foreign languages is in contrast to that of the academic experts, who assigned it a score of just 13.3%. The regular use of a second language is becoming more prevalent in firms operating abroad, a factor that explains why representatives of the labour market value a second language more highly than do academics.

The frequent use of technical-specific competencies was a crucial indicator (19.1%), especially for labour experts (25.4% vs. 17.8% for academics). Both groups assigned a high score (16.0%) to *forma mentis* or problem setting (Table 3). Academics considered that the time elapsed between recruitment and first promotion strongly depends on technical-specific competencies (14.6%), whereas labour experts gave this a low score (4.5%). In this respect, the academic point of view can be seen to be more idealistic than that of the labour market experts who want graduates to possess a more technical and production-oriented background than they actually do.

Soft skills can be strengthened at university, but they are mainly dependent upon personal experience; according to the labour representatives, the attitude to motivate

colleagues is derived first from soft skills (30.8%). A readiness to deal with new problems and to use new tools is also associated with motivation (25.2%), as far as the attitude to teach fellow workers and to support young colleagues (26.2%). Academics also value these skills, although not as strongly as labour experts.

Relational skills place the graduate at the core of a group, connected in turn to other units of the economic system. The graduate is deemed to be successful if he/she can act as an intermediary between opposing interests (17.9%), argue his/her own ideas in public (16.9%) and deal with customers and/or suppliers (16.0%).

Some indicators of organisational skills are similar to those associated with life-long learning and communication skills. The number of soft skills components of the study could perhaps be reduced in a future exercise.

A work-oriented personality was seen by all as someone who is motivated and finds the results of work stimulating (25.5%). Labour experts in particular scored these traits high (33%). The relevance of motivation in different components of GHC indicates that this characteristic could be isolated in a single sub-dimension. An additional interesting indicator of a work-oriented personality was the graduate's availability to work in conditions involving personal costs, such as working after hours or during the weekend (17.7%). Both groups of experts placed a comparable value on precision, accuracy and diligence (about 17%).

The final and most discretionary dimension of GHC is the inclination for individuals to take social ethics into account. It is debatable whether this dimension can be strengthened during university. Some may argue that universities are unable to convey a value system to students; others may contend that it is inappropriate for universities to take on such a role as other social institutions are charged with the development of individual and social ethics. Nevertheless, both academic and labour experts agree on the relative importance of social ethics and its indicators, the most relevant of which are the propensity to give credit to other people at work (24.0%), compliance with internal rules and the law (21.6%), and social equity (20.5%).

4 Concluding Remarks

Based on our results, we conclude that the data collection method worked well. A good proportion of experts completed all the three questionnaires and the responses that were collected were consistent among experts. Dimension and indicator selection and weight estimation could be performed thoroughly, essentially at no cost. The method may, therefore, be used to elicit ideas and preferences relating to hierarchical dimensions of GHC and to quantify their weights.

The whole procedure took just 50 days. The promptness of experts in answering complex questionnaires might be explained by the relevance of the issue, institutional position and contact mode. It might be possible to repeat the exercise in a shorter time, provided the concept to be measured is simple and motivating.

The consistency found in the responses stems, no doubt, from a commonality in opinions on the topic. We suggest that the indirect estimation procedure of weights – the analysis of the dominance matrix – is the most convenient way of obtaining

responses. This approach overcomes potential consequences that could occur as a result of using a numerical system in direct preference elicitation.

One limitation of the study was the difficulty that experts encountered in guessing the indicators pertinent to each GHC dimension. Some stated that they were unable to decipher the meaning of a dimension's indicator; the majority wrote that indicator guessing was the most difficult aspect. The dimension definitions and the classification of indicators could be restructured in future studies. The methodology for efficiently defining measurable indicators remains a matter for debate.

Any HC measurement structure is "situated" or embedded in a given situation. In this instance, the study focuses on outcomes of new graduates in Italy. Our HC structure may be adopted for measuring external educational effectiveness since the Delphi questions were openly posed for defining a *GHC targeted to effectiveness representation* and the stakeholders actively participated in that definitional process.

One question that remains unanswered is whether dimensions other than those examined, in particular social dimensions that differ from labour, should be considered in assessing the effectiveness of university education.

References

Becker, G. S. (1993). *Human capital*. New York: Columbia University Press.

Coleman, J. S. (1990). *Foundations of social theory*. London: The Belknap Press of Harvard University Press.

Fabbris, L. (2010). Dimensionality of scores obtained with a paired-comparison tournament system of questionnaire items. In F. Palumbo, C. N. Lauro, & M. J. Greenacre (Eds.), *Data analysis and classification. Proceedings of the 6th conference of the classification and data analysis group of the Società Italiana di Statistica* (pp. 115–162). Berlin-Heidelberg: Springer-Verlag.

Fabbris, L., D'Ovidio, F. D., Pacinelli, A., & Vanin, C. (2008). Profili professionali di addetti alle risorse umane sulla base di due panel giustapposti di esperti Delphi-Shang. In L. Fabbris (Ed.), *Definire figure professionali tramite testimoni privilegiati* (pp. 101–134). Padova: Cleup.

Ford, D. (1975). Shang inquiry as an alternative to Delphi: Some experimental findings. *Technological Forecasting and Social Change, 7*(2), 139–164.

Hanushek, E. A. (1997). Assessing the effects of school resources on economic performance. *Education Evaluation and Policy Analysis, 19*(2), 141–164.

OECD (1998). *Human capital investment: an international comparison*. Paris: OECD.

Saaty, T. L. (1983). Rank according to Perron: A new insight. *Mathematics Magazine, 60*(4), 211–213.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon.

Using Poset Theory to Compare Fuzzy Multidimensional Material Deprivation Across Regions

Marco Fattore, Rainer Brüggemann, and Jan Owsiński

Abstract In this paper, a new approach to the fuzzy analysis of multidimensional material deprivation data is provided, based on partial order theory. The main feature of the methodology is that the information needed for the deprivation assessment is extracted directly from the relational structure of the dataset, avoiding any kind of scaling and aggregation procedure, so as to respect the ordinal nature of the data. An example based on real data is worked out, pertaining to material deprivation in Italy for the year 2004.

1 Introduction

The aim of this paper is to present new tools for fuzzy analysis of multidimensional material deprivation data and, more generally, for the analysis of multivariate ordinal datasets for evaluation and ranking purposes. The methodology combines fuzzy set theory and partial order theory and can be applied to evaluation problems in several fields, such as assessing quality of life, quality of services or quality of the environment (Brüggemann et al. 2001) to mention a few. The main feature of the approach is that the ordinal nature of the data is fully respected, avoiding any kind of scaling and aggregation procedure. The information needed for the evaluation process is extracted directly from the relational structure of the dataset, without turning ordinal scores into cardinal numbers. This is accomplished by means of partial order theory, a set of algebraic tools that provides the right formal language to tackle ordinal evaluation problems. For the sake of clarity and readability, all the necessary algebraic tools will be presented through a leading example pertaining to material deprivation in Italian macro-regions.

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_6,

2 Material Deprivation Data

Data about household material deprivation in Italian macro-regions are extracted from the EU-SILC survey, for the year 2004. To keep computations simple, only five variables have been considered, namely:

1. HS040 - *Capacity to afford paying for one week annual holiday away from home*;
2. HS050 - *Capacity to afford a meal with meat, chicken, fish (or vegetarian equivalent) every second day*;
3. HS070 - *Owning a phone (or mobile phone)*;
4. HS080 - *Owning a colour TV*;
5. HS100 - *Owning a washing machine*.

All five variables are coded in a binary form: 0 if the household is not deprived on the item and 1 if it is deprived. Each possible sequence of five binary digits defines a different *deprivation state*; for instance, the state $s = 10011$ stands for deprivation on HS040, HS080 and HS100 and non-deprivation on HS050 and HS070. Deprivation states can be ordered in a natural way, according to the following definition:

Definition 2.1 *Let s and t be two material deprivation states. We will write $s \trianglelefteq t$ if and only if $s_i \leq t_i \ \forall\, i = 1, \ldots, 5$, where s_i and t_i are the $i-th$ digits of the binary representations of s and t respectively. State s is (strictly) less deprived than state t ($s \triangleleft t$) if and only if $s \trianglelefteq t$ and there exists at least one j such that $s_j < t_j$.*

Clearly, not all the deprivation states can be ordered, based on the previous definition, since there may be incomparabilities among them (e.g. consider states 10000 and 00001). As a result, the set of deprivation states gives rise to a *partially ordered set* (or *poset*, for short). Formally, a poset is a set equipped with a partial order relation, that is a binary relation satisfying the properties of *reflexivity*, *antisymmetry* and *transitivity* (Davey and Priestley 2002). A finite poset P (i.e. a poset defined on a finite set) can be easily depicted by means of a *Hasse diagram*, which is a particular kind of directed graph, drawn according to the following two rules: (1) if $s \triangleleft t$, then node t is placed above node s; (2) if $s \triangleleft t$ and there is no other state w such that $s \triangleleft w \triangleleft t$ (i.e. if t *covers* s), then an edge is inserted linking node t to node s. By transitivity, $s \triangleleft t$ in P, if and only if in the Hasse diagram there is a descending path linking the corresponding nodes; otherwise, states s and t are *incomparable* ($s \,||\, t$). Since in the example five binary variables are considered, the poset L of all possible deprivation states is composed of 32 nodes and 27 of them are actually observed in the data pertaining to Italy (referring to a sample of 24,202 households). The Hasse diagram of L is shown in Fig. 1 (observed states are represented as black nodes). The top node ($\top$) and the bottom node ($\bot$) represent the *completely deprived* state (11111) and the *completely non-deprived* state (00000) respectively.

Even if some possible deprivation states are not realized in the observed population, they have a clear meaning and could indeed be realized in other circumstances (e.g., by different populations or by the same population in different times). So, to get robust results and be able to compare population over time or space, we address

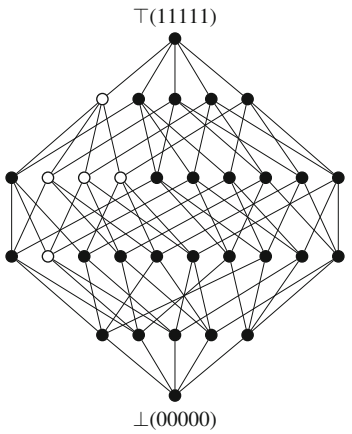


Fig. 1 Hasse diagram for the poset of material deprivation states (black nodes refer to states observed in Italy)

all of the states of L, rather than only the realised ones. Apparently, the partial order induced by Definition 2.1 is poorly informative about the deprivation level of many poset states. For example, states 10000 and 01111 are considered as incomparable, even if the latter seems to be much more deprived than the former. Nevertheless, as will be shown in the following, the different deprivation level of such states is clearly revealed when their different position in the Hasse diagram is considered and when the global relational pattern of the data is analyzed by means of partial order tools.

3 The Material Deprivation Membership Function

The central problem in assessing material deprivation is to assign a degree of deprivation to each state in L, that is to compute the *material deprivation membership function* for each deprivation state. Formally, a deprivation membership function is an *order preserving* map $m(\cdot)$ from L to $[0, 1]$, that is a map

$$\begin{aligned} m : L &\mapsto [0, 1] \\ : s &\to m(s) \end{aligned} \tag{1}$$

such that

$$s \trianglelefteq t \Rightarrow m(s) \leq m(t). \tag{2}$$

Clearly there is no unique criterion to select a deprivation membership function, out of such a class of maps. In this paper, we follow the point of view of 'response structure of the population', as in (Cerioli and Zani 1990) and assume that the degree of deprivation, assigned by $m(\cdot)$ to each state in L, depends upon the combined

different assessments given by a population of 'judges'. The first step in view of the definition of $m(\cdot)$ is therefore to make explicit how the set of judges is built.

Linear extensions of a poset. The key idea to identify a suitable set of judges can be explained as follows. Judges produce rankings of deprivation states out of the poset L; when accomplishing this task, they are free to order incomparable pairs as preferred (no ties are allowed), but they cannot violate the constraints given by the original partial order, i.e. if $s \lhd t$ in L, then any judge must rank t above s in his own deprivation ranking. Thus, the set of all possible different judges (i.e. judges not producing the same rankings) coincides with the set of all the *linear extensions* of L. A linear extension of a poset P is a linear ordering of the elements of P which is consistent with the constraints given by the partial order relation. For example, if P is composed of three elements x, y and z, with $y \leq x$, $z \leq x$ and $y \,||\, z$, only two linear extensions are possible, namely $z \leq y \leq x$ and $y \leq z \leq x$, since x is greater than both y and z in P. The set of all the linear extensions of a poset P is denoted by $\Omega(P)$; it comprises all the linear orders compatible with P and identifies uniquely the partial order structure (Neggers and Kim 1998).

Up-sets, down-sets and the deprivation border. In view of a fuzzy assessment of material deprivation, three relevant subsets of L can be identified, namely

- the set D of *certainly deprived* states: $D = \{s \in L : m(s) = 1\}$;
- the set W of *certainly non-deprived* states: $W = \{s \in L : m(s) = 0\}$;
- the set A of *ambiguously deprived* states: $A = \{s \in L : 0 < m(s) < 1\}$.

According to (1) and (2), if $s \in D$ and $s \trianglelefteq t$, then $t \in D$; similarly, if $s \in W$ and $t \trianglelefteq s$, then $t \in W$. In poset theoretical terms, sets like D and W are called *up-sets* and *down-sets*, respectively. When fuzzy poverty is assessed in monetary terms, a threshold τ is usually identified separating certainly poor people from the rest of the population. A similar threshold can also be defined for the poset of material deprivation states. Given the up-set D, there is a unique subset $\underline{d} \subseteq D$ of mutually incomparable elements (a so called *antichain*), such that $s \in D$ if and only $d \trianglelefteq s$ for some $d \in \underline{d}$ (Davey and Priestley 2002). The up-set D is said to be generated by $\underline{d}$ (in formulas, $D = \uparrow\underline{d}$). Excluding the trivial cases of $D = L$ and $D = \top$, any element of the generating antichain is covered only by elements of D and covers only elements of $L \backslash D$, so that it shares the same role that τ has in the monetary case. For this reason, $\underline{d}$ can be called the *material deprivation border*.

Membership function definition. To define the membership function, we need (i) to determine how linear extensions (i.e. the 'judges') $\omega \in \Omega(L)$ assign the degrees of deprivation $Dep_\omega(s)$ to a state $s \in L$ and (b) to decide how to combine all such degrees into the 'final' degree $m(s)$. First of all, let us assume that an antichain $\underline{d}^*$ is chosen as the deprivation border (this means that all judges agree to assign degree of deprivation one to all elements in $\uparrow\underline{d}^*$). At a purely illustrative level[1] we can put

[1] The choice of a meaningful deprivation border is subjective and requires experts judgments. In this methodological paper, we do not deal with this fundamental issue and take the border as given.

$$\underline{d}^* = (10100, 11000). \tag{3}$$

(i) Once the border has been identified, the simplest way we can define the degree of deprivation of a state s in a linear extension ω is to put

$$Dep_\omega(s) = \begin{cases} 1 & if\ \exists\, d \in \underline{d}^* \text{such that } d \leq s \text{ in } \omega; \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Explicitly, ω assigns degree of deprivation 1 to each element of $\underline{d}^*$ and to each element of L that is ranked by the ω above *at least one* element of $\underline{d}^*$; conversely, ω assigns the degree of deprivation 0 to those states that are ranked, in ω, below *all* the elements of $\underline{d}^*$.

(ii) The membership function for state s (that we write as $m(s|\,\underline{d}^*)$ since it depends upon the choice of the material deprivation border) is then computed as

$$m(s|\,\underline{d}^*) = \frac{|\{\omega : Dep_\omega(s) = 1\}|}{|\Omega(L)|}, \tag{5}$$

i.e. as the fraction of linear extensions ranking state s as deprived, given $\underline{d}^*$.
Before turning to the problem of the computation of $m(s|\,\underline{d}^*)$, some comments to definitions (4) and (5) are in order.

1. Each linear extension classifies the states of L in binary terms; this means that the fuzziness we are dealing with is due to different responses by different judges, while the single judge acts in a crisp way.
2. Taking into account all linear extensions, our measure of fuzziness takes care of both comparable and not comparable, i.e. non commensurable, deprivation states.
3. Given the deprivation border (3), one easily sees that the set D is composed of the states 10100, 10101, 10110, 10111, 11000,11001, 11010, 11011, 11100, 11101, 11110, 11111. In fact, each of these states belongs to the up-set of an element of the deprivation border $\underline{d}^*$ (for example, state 10101 is in the up-set of 10100), so that each linear extension assigns to them the deprivation degree of 1.
4. Formula (5) implies that the set of certainly non-deprived states is composed of all those states that are less deprived than any element of $\underline{d}^*$. In our specific example, it turns out that $W = \{00000, 10000\}$.

Membership function computation. To compute the fraction of linear extensions assigning degree of deprivation 1 to a state s, we will have to list all the elements of $\Omega(L)$ and select those where s is ranked above d for some $d \in \underline{d}^*$. Unfortunately, listing all the linear extensions of a poset is computationally impossible (unless the poset is very small or contains very few incomparabilities), so mutual ranking frequencies must be estimated, based on a sample of linear extensions. The computations presented in this paper are performed by running the Bubley-Dyer

Table 1 Membership function $m(s|10100, 11000)$

State s:	00000	00001	00010	00011	00100	00101	00110	00111
$m(s\|10100, 11000)$	0.00	0.11	0.11	0.65	0.06	0.66	0.66	0.98
State s	01000	01001	01010	01011	01100	01101	01110	01111
$m(s\|10100, 11000)$	0.06	0.66	0.66	0.98	0.67	0.98	0.98	1.00
State s	10000	10001	10010	10011	10100	10101	10110	10111
$m(s\|10100, 11000)$	0.00	0.67	0.67	0.98	1.00	1.00	1.00	1.00
State s	11000	11001	11010	11011	11100	11101	11110	11111
$m(s\|10100, 11000)$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

algorithm, which is the most efficient known algorithm for (quasi) uniform sampling of linear extensions (Bubley and Dyer 1999). Particularly, the membership function $m(s\,|10100, 11000)$ has been estimated using a sample of 10^8 linear extensions and is reported in Table 1.

It must be noted that the table actually reports an estimation of the true values of the membership function; this explains why $m(01111\,|10100, 11000) = 1$, even if state 01111 does not belong to D.

4 Fuzzy Material Deprivation in Italian Macro-Regions

The membership function computed in the previous paragraph has been used to compute the amount of fuzzy poverty in Italy and in five Italian macro-regions, namely North–West, North–East, Centre, South and Islands, based on the data described at the beginning of the paper. Two normalized indicators have been considered. The first indicator FH is the fuzzy extension of the classical Head Count Ratio H; it is defined as

$$FH = \frac{\sum_{s\in L} m(s \mid \underline{d}^*) \cdot |s|}{\sum_{s\in L} |s|} \tag{6}$$

where $|s|$ is the number of people occupying state s. FH is simply the ratio between the fuzzy cardinality of the set of people having a non-null degree of deprivation and the cardinality of the entire population. The second indicator C is defined as

$$C = \frac{\sum_{s\in L} m(s|\underline{d}^*) \cdot |s|}{\sum_{s\in L} \delta(s) \cdot |s|}, \quad \delta(s) = \begin{cases} 1 & \text{if } m(s|\underline{d}^*) > 0 \\ 0 & \text{if } m(s|\underline{d}^*) = 0. \end{cases} \tag{7}$$

If $\sum_{s\in L} \delta(s) \cdot |s| = 0$, C is set to 0. C measures the mean level of 'deprivation certainty' characterizing the subpopulation of people occupying states with a non-null degree of deprivation. The results are reported in Table 2.

Although the present application has an illustrative goal, the results obtained are consistent with the territorial differences in Italy. Southern regions show a much

Table 2 Values of FH and C for Italian macro-regions and the whole country, year 2004

	FH	C
North-West	0.04	0.76
North-East	0.05	0.76
Centre	0.06	0.88
South	0.15	0.93
Islands	0.15	0.94
Italy	0.07	0.86

greater incidence of material deprivation than regions in the North and in the Centre. Also index C is markedly higher in the Southern regions, revealing a possible social polarization: not only is material deprivation greater in the Southern part of Italy, but in those regions deprived people are also 'almost' certainly deprived, i.e. in a fuzzy sense they belong definitely to the set of deprived people.

5 Conclusion

In this paper, we have shown how poset theory provides an effective setting for fuzzy modeling of multidimensional material deprivation data and, more generally, of multidimensional ordinal datasets. The main advantage of the methodology is that, differently from other approaches (Cerioli and Zani 1990, Lemmi and Betti 2006), it relies only on the ordinal nature of the data, without supposing any quantitative model behind them. This way, a sound numerical evaluation of the deprivation degrees is computed out of qualitative ordinal information, preserving those ambiguities that are co-essential to the concepts that are dealt with. This is consistent with Sen's point of view, that if there is some ambiguity in a concept, 'a precise representation of that ambiguous concept must preserve that ambiguity' (Sen 1992). Here, partial order theory plays a role similar to that of linear algebra in quantitative multivariate data analysis: it makes it possible to represent and exploit the structure of the data and to extract information directly out of it. There are indeed some open issues. The combinatoric approach used in the paper is in fact particularly suitable when the number of states is not too large. To extend the proposed methodology to situations where many variables or states are considered, different approaches are being developed. They combine (1) techniques for the clustering of states, based on the algebraic tools of congruences (Cheung and Vogel 2005); (2) the development of better performing software procedures to shorten the computation time in the estimation of mutual ranking frequencies; (3) the identification of analytical formulas yielding approximated values of such frequencies, directly out of the poset topology. In conclusion, partial order theory paves the way to a new approach for studying multidimensional systems of ordinal data; at the same time, it calls for further methodological research, so as to extend and tune partial order concepts and techniques towards the needs of applied statistical analysis and modeling.

References

Brüggemann, R., Halfon, E., Welzl, G., Voigt, K., & Steinberg C. (2001). Applying the concept of partially ordered sets on the ranking of near-shore sediments by a battery of tests. *Journal of Chemical Information and Computer Science*, *41*, 918–925.

Bubley, R., & Dyer, M. (1999). Faster random generation of linear extensions. *Discrete mathematics*, *201*, 81–88.

Cerioli, A., & Zani, S. (1990). A fuzzy approach to the measurement of poverty. In C. Dagum, & M. Zenga (Eds.), *Income and wealth distribution, inequality and poverty* (pp. 272–284). Berlino Heidelberg: Springer-Verlag.

Cheung, K. S. K., & Vogel, D. (2005). Complexity reduction in lattice-based information retrieval. *Information Retrieval, 8*, 285–299.

Davey, B. A., & Priestley B. H. (2002). *Introduction to lattices and order*. Cambridge: Cambridge University Press.

Lemmi, A., & Betti, G. (2006). *Fuzzy set approach to multidimensional poverty measurement*. New York: Springer.

Neggers, J., & Kim, H. S. (1988). *Basic posets*. Singapore: World Scientific.

Sen, A. K. (1992). *Inequality reexamined*. Oxford: Clarendon Press.

Some Notes on the Applicability of Cluster-Weighted Modeling in Effectiveness Studies

Simona C. Minotti

Abstract In the nineties, numerous authors proposed the use of Multilevel Models in effectiveness studies. However, this approach has been strongly criticized. Cluster-Weighted Modeling (CWM) is a flexible statistical framework, which is based on weighted combinations of local models. While Multilevel Models provide rankings of the institutions, in the CWM approach many models of effectiveness are estimated, each of them being valid for a certain subpopulation of users.

1 Introduction

In the nineties, a number of authors proposed the use of Multilevel Models (Goldstein 1995) in the context of effectiveness studies, given the typical multilevel structure of the data (students clustered in schools, patients grouped in hospitals, etc.). We refer, here, to the comparisons of institutional performance in public sector activities such as education, health and social services, i.e., the well-known rankings (see Bryk and Raudenbush 2002, Chap. 5). In educational studies, the effectiveness indicates the added value in students' achievement level produced by schools, while in healthcare studies, it refers to the effect of hospital care on patients.

Multilevel Models may produce, however, not reliable rankings (Goldstein and Spiegelhalter 1996); moreover, in the case of large datasets (like in regional or national studies based on administrative data), significance tests of the parameters of linear models always lead to the rejection of the null hypothesis (Vroman Battle and Rakow 1993).

In order to overcome the drawbacks of Multilevel Models in the context of effectiveness studies, we propose the use of Cluster-Weighted Modeling (CWM) (Gershenfeld et al. 1999).

CWM is a flexible statistical framework for capturing local behaviour in heterogenous populations, which is based on weighted combinations of local models. Developed by Gershenfeld, Schoener and Metois in order to recreate a digital violin with traditional inputs and realistic sound (Gershenfeld et al. 1999), CWM has been set in a statistical framework by Ingrassia et al. (2009).

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_7,

In the context of effectiveness studies, while Multilevel Models provide rankings of the institutions, in the CWM approach many models of effectiveness are estimated, each of them being valid for a certain subpopulation of users.

Some weak points of Multilevel Models in effectiveness studies will be highlighted in Sect. 2; CWM will be introduced in Sect. 3; the applicability of CWM in effectiveness studies will be discussed in Sect. 4; in Sect. 5 we will provide conclusions and discuss further research.

2 Some Weak Points of Multilevel Models in Effectiveness Studies

We refer, here, to a Two-Level Random Intercept Model; following the setting-out in Goldstein (1995), it may be written as:

$$y_{ij} = \beta_0 + \sum_{g=1}^{G} \beta_g x_{g\,ij} + u_j + e_{ij}, \tag{1}$$

where, in the context of effectiveness studies (see Bryk and Raudenbush 2002), y_{ij} is the outcome measured on the i-th individual belonging to the j-th institution ($i = 1, \ldots, n_j$; $j = 1, \ldots, Q$; $N = n_1 + \ldots + n_j + \ldots + n_Q$); β_0 is the average outcome across all individuals and all institutions; β_g is the effect of the individual-specific covariate X_g ($g = 1, \ldots, G$); u_j is the second-level random effect associated with the j-th institution and is known as the effectiveness parameter, i.e., the effect of the j-th institution on the outcome y_{ij}, adjusted for individual-level characteristics ($u_j \sim N(0, \tau_{00})$, where τ_{00} is the variation in intercepts between institutions); e_{ij} is the first-level random effect associated with the i-th individual belonging to the j-th institution ($e_{ij} \sim N(0, \sigma^2)$, where σ^2 is the variation within institutions).

The plot of confidence intervals for ordered effectiveness parameters provides a ranking of the institutions, adjusted by individual-specific characteristics.

However, there are situations where Multilevel Models appear to be not informative and prevent to distinguish among the institutions. In fact, this approach may produce not reliable rankings (i.e., characterized by large confidence intervals for the effectiveness parameters). In this case, "an overinterpretation of a set of rankings where there are large uncertainty intervals can lead both to unfairness and inefficiency and unwarranted conclusions about changes in ranks" (Goldstein and Spiegelhalter 1996). Then, "league tables of outcomes are not a valid instrument for day-to-day performance management by external agencies" (Lilford et al. 2004). Large confidence intervals may be a consequence of both heterogeneity of individual-level relationships and heterogeneity of the individuals within institutions, like in regional or national studies (see Fig. 1).

Moreover, in the case of large data, significance tests for the parameters of linear models always lead to the rejection of the null hypothesis (Vroman Battle and

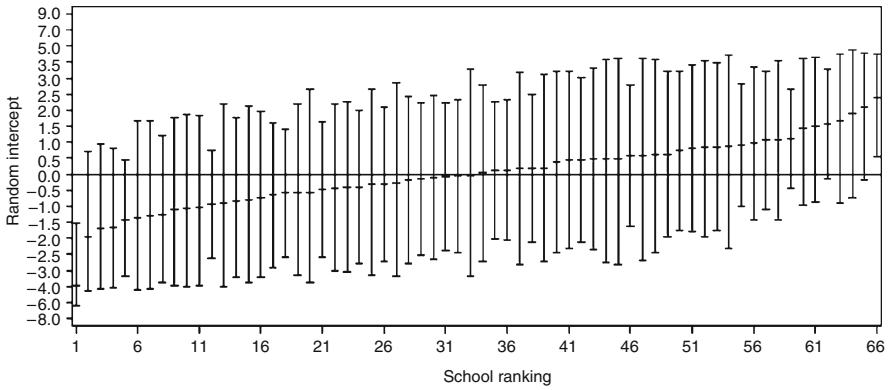


Fig. 1 Example of a not reliable ranking of schools

Rakow 1993). In these cases, a single global model is not sufficient and methods able to capture local behaviour seem necessary. A first attempt in this direction is described in Minotti and Vittadini (2010).

3 Cluster-Weighted Modeling

CWM is a flexible statistical framework for capturing local behaviour in heterogenous populations, based on joint probability $p(\mathbf{x}, y)$ estimated from a set of pairs of input–output learning data $\{(\mathbf{x}_n, y_n)\}_{n=1,\ldots,N}$.

Let $(\mathbf{X}, Y)$ be a pair of a random vector $\mathbf{X}$ and a random variable Y defined on Ω with joint probability distribution $p(\mathbf{x}, y)$, where $\mathbf{X}$ is the d-dimensional input vector with values in some space $\mathscr{X} \subseteq \Re^d$ and Y is a response variable having values in $\mathscr{Y} \subseteq \Re$. Assume that Ω can be partitioned into G disjoint groups, say $\Omega_1 \cup \cdots \cup \Omega_G$.

CWM decomposes the joint probability $p(\mathbf{x}, y)$ as

$$p(\mathbf{x}, y) = \sum_{g=1}^{G} p(y|\mathbf{x}, \Omega_g)\, p(\mathbf{x}|\Omega_g)\, \pi_g, \tag{2}$$

where π_g is the mixing weight of the group Ω_g, $p(\mathbf{x}|\Omega_g)$ is the probability density of $\mathbf{x}$ given Ω_g and $p(y|\mathbf{x}, \Omega_g)$ is the conditional density of the response variable Y given the predictor vector $\mathbf{x}$ and the group Ω_g. The $p(\mathbf{x}|\Omega_g)$, $(g = 1, \ldots, G)$, in (2) are usually assumed to be multivariate Gaussians, that is $p(\mathbf{x}|\Omega_g) = \phi_d(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$. Moreover, the $p(y|\mathbf{x}, \Omega_g)$, $(g = 1, \ldots, G)$, can be modeled again by Gaussian distributions with variance $\sigma^2_{\varepsilon,g}$ around some local models, here restricted to the case $\gamma_g(\mathbf{x}) = \mathbf{b}'_g\mathbf{x} + b_{g0}$, with $\mathbf{b}_g \in \Re^d, b_{g0} \in \Re$

and then

$$p(\mathbf{x}, y; \boldsymbol{\theta}) = \sum_{g=1}^{G} \phi(y; \mathbf{b}_g' \mathbf{x} + b_{g0}, \sigma_{\varepsilon,g}^2)\, \phi_d(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\, \pi_g. \tag{3}$$

The estimation of the parameters can be performed by means of the EM algorithm (see Ingrassia et al. 2009); the number of groups may be chosen on the basis of the Bayesian Information Criterion, like in traditional Mixture Models (Frühwirth-Schnatter 2005).

Recently, Ingrassia et al. (2009) have demonstrated that the Gaussian CWM is a generalization of Finite Mixtures of Regression (FMR) (Frühwirth-Schnatter 2005)

$$f(y|\mathbf{x}; \boldsymbol{\psi}) = \sum_{g=1}^{G} \phi(y; \mathbf{b}_g' \mathbf{x} + b_{g0}, \sigma_{\epsilon,g}^2)\pi_g \tag{4}$$

and Finite Mixtures of Regression with Concomitant Variables (FMRC) (Dayton and Macready 1988)

$$f^*(y|\mathbf{x}; \boldsymbol{\psi}^*) = \sum_{g=1}^{G} \phi(y; \mathbf{b}_g' \mathbf{x} + b_{g0}, \sigma_{\epsilon,g}^2) p(\Omega_g|\mathbf{x}, \boldsymbol{\xi}), \tag{5}$$

where the mixing weight $p(\Omega_g|\mathbf{x}, \boldsymbol{\xi})$ is modeled by a multinomial logit.

4 Applicability of CWM in Effectiveness Studies

In order to overcome the drawbacks of Multilevel Models in the context of effectiveness studies, we propose the use of CWM.

For sake of brevity, the applicability of CWM will be illustrated by means of two artificial examples in the field of accountability systems, which typically aim to explain and predict students' achievement outcomes. In particular, we refer to the study of the relationship between Exit achievement (assumed here as a continuous outcome) and Intake achievement (assumed here as a continuous predictor) measured on some hypothetical students belonging to some hypothetical schools.

The examples try to reproduce some typical situations in effectiveness studies and aim to show when CWM is more informative than Multilevel Modeling.

Example 1. The first example regards the simulation of the heterogeneity of student-level relationship among schools by means of two linear effectiveness models with equal distribution for the Intake achievement (i.e., with homogenous students) and which differ in intercept, only.

The sample was generated according to the following parameters:
Mod. 1 (Sch.1,2):

$N_1 = 100$, $p(x|\Omega_1) = \phi(x; 5, 0.2)$, $p(y|x, \Omega_1) = \phi(y; 2 + 6x, 0.2)$
Mod. 2 (Sch.3,4):

$N_2 = 100$, $p(x|\Omega_2) = \phi(x; 5, 0.2)$, $p(y|x, \Omega_2) = \phi(y; 8 + 6x, 0.2)$.

The scatter plot of the original data (labeled by school) is reported in the first panel of Fig. 3 and shows that students from School 3 and 4 have higher Exit achievement than students from School 1 and 2. The ranking of the four schools obtained by means of a Two-Level Random Intercept Model is reported (see Fig. 2) and shows that School 3 and 4 are more effective than School 1 and 2.

In Fig. 3, the reclassified data by means of FMR, FMRC and CWM respectively, are also reported. The criterion is that a student is classified into the local effectiveness model with the maximum posterior probability.

In this special case, the three competitive models are equivalent and provide analogous information with respect to Multilevel Modeling. In fact, both FMR, FMRC

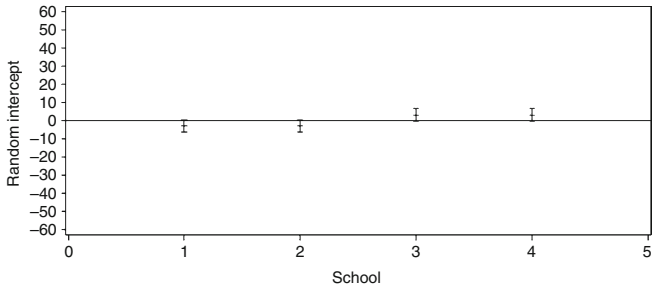


Fig. 2 Example 1. Ranking of the schools

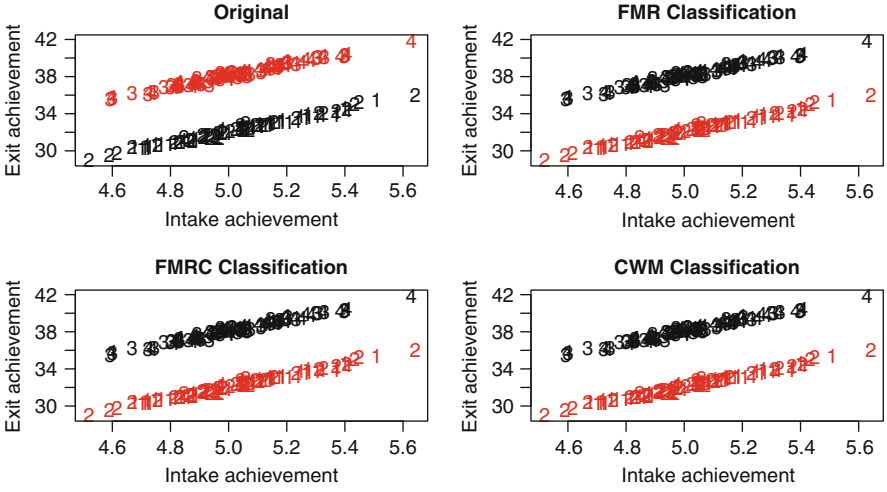


Fig. 3 Example 1. Original and reclassified data (FMR, FMRC, CWM)

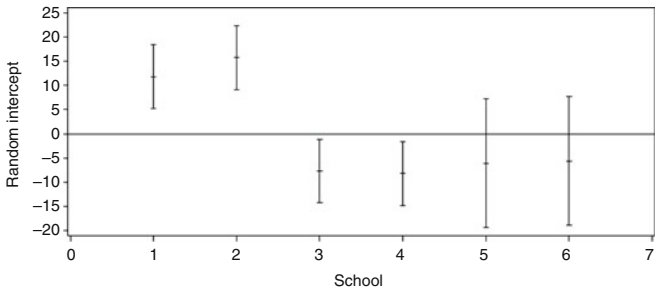


Fig. 4 Example 2. Ranking of the schools

and CWM indicate that students from School 3 and 4 have a better performance than students from School 1 and 2.

Example 2. The second example regards a more complex situation, i.e., the simultaneous heterogeneity of student-level relationship among schools and heterogeneity of the students within schools. We have three linear effectiveness models with different distribution for the Intake achievement (i.e., not homogenous students).

The sample was generated according to the following parameters:
Mod. 1 (Sch.1,2):

$N_1 = 100$, $p(x|\Omega_1) = \phi(x; 5, 2)$, $p(y|x, \Omega_1) = \phi(y; 6 + 40x, 2)$

Mod. 2 (Sch.3,4):

$N_2 = 600$, $p(x|\Omega_2) = \phi(x; 10, 2)$, $p(y|x, \Omega_2) = \phi(y; -1.5 + 40x, 2)$

Mod. 3 (Sch.5,6):

$N_3 = 300$, $p(x|\Omega_3) = \phi(x; 20, 2)$, $p(y|x, \Omega_3) = \phi(y; -7 + 150x, 2)$.

The scatter plot of the original data (labeled by school) is reported in the first panel of Fig. 5. The ranking of the six schools obtained by means of a Two-Level Random Intercept Model is reported (see Fig. 4). A Two-Level Random Coefficient Model has been also estimated; the random effects, however, were not statistically significant.

This is an example where Multilevel Modeling is not particularly informative, due to the large confidence intervals. In fact, the ranking is not reliable, because School 5 and 6 are not clearly distinguishable from School 3 and 4. In this case, methods able to model local behaviour enable to explain these differences.

From a methodological point of view, CWM only is able to correctly classify the original observations (see Fig. 5); the reason is given by the different decision surfaces characterizing the three competitive models (see for more details Ingrassia et al. (2009)). From an interpretative point of view, CWM is much more informative than Multilevel Modeling. In fact, it results that School 1 and 2 teach to students with an Intake achievement which is lower than students from other schools (especially those from School 5 and 6), but allow them to reach the highest levels of Exit achievement; consequently they are more effective than other schools. Instead, Schools 3 and 4 teach to students with a medium level of Intake achievement, but they follow an effectiveness model which is characterized by a weak negative relationship between Intake and Exit achievement. Then, Schools 5 and 6 teach to

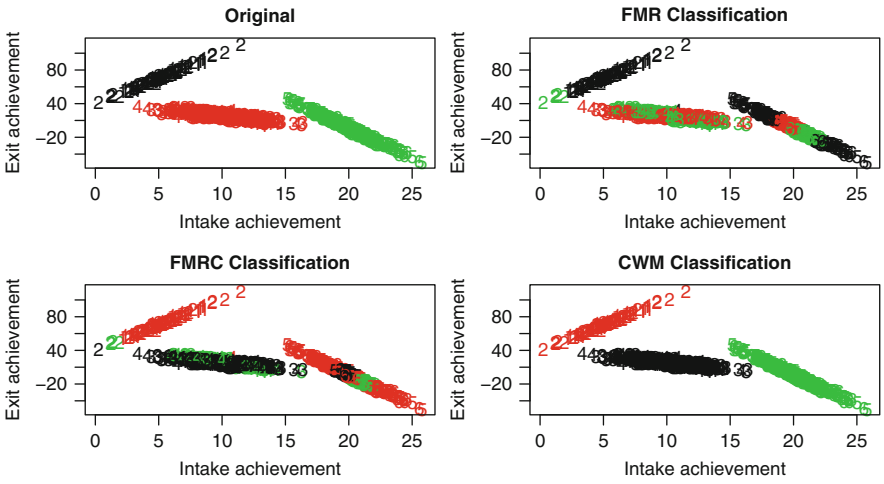


Fig. 5 Example 2. Original and reclassified data (FMR, FMRC, CWM)

students with the highest level of Intake achievement, but they follow an effectiveness model which is characterized by a strong negative slope; consequently, they are the worse schools.

The examples enable some general considerations about the applicability of CWM in the context of effectiveness studies.

Suppose to have many schools with not homogenous students, described by a set of explicative variables. CWM estimates many local models which express the local relationship between the Exit achievement and the characteristics of the students, regardless of the multilevel structure (i.e., the membership of a student to a school). Each student is classified into the local model with the maximum posterior probability. From students it is possible to identify schools and, consequently, which effectiveness models follow the schools with respect to specific subgroups of students. This enables analysis of the effectiveness levels for each school (e.g., a school could be clever to teach to students with low levels of Intake achievement and not clever with students characterized by high levels of Intake achievement), and comparisons among institutions with similar behaviour (e.g., by means of Multilevel Modeling enriched by school-level variables). Moreover, in terms of policies, this means having an instrument to address homogenous subpopulations of students to the most effective schools.

5 Conclusions and Further Research

In this paper we have proposed CWM as a new methodology in the context of effectiveness studies and a useful tool to support public policy decisions. In particular, this approach appears to be particularly suitable for statistical modeling of

administrative data (large data), where the high complexity may not be interpreted by the multilevel structure only.

Further research will regard the extension of Cluster-Weighted Modeling in the following three directions: (1) mixed variables, given that typical outcomes and predictors involved in effectiveness studies are not exclusively real-valued (e.g., achievement is an ordinal variable; mortality rate is a dichotomous variable); (2) non-linear local models, given that the local relationships between outcomes and predictors may be non linear; (3) multilevel data structures, as proposed in Galimberti and Soffritti (2007) for Mixture Models, in order to allow some of the parameters of the conditional densities to differ across second-level units (schools, hospitals, etc.).

References

Bryk, A. S., & Raudenbush, S. W. (2002). Hierarchical linear models. *Applications and data analysis methods*. Newbury Park, CA: Sage Publications.

Dayton, C. M., & Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association*, *83*, 173–178.

Frühwirth-Schnatter, S. (2005). *Finite mixture and markov switching models*. Heidelberg: Springer.

Galimberti, G., & Soffritti, G. (2007). Multiple cluster structures and mixture models: Recent developments for multilevel data. In: *Book of short papers CLADAG 2007*. Macerata: EUM.

Gershenfeld, N., Schöner, B., & Metois, E. (1999). Cluster-weighted modeling for time-series analysis. *Nature*, *397*, 329–332.

Goldstein, H. (1995). *Multilevel statistical models*. London: Arnold.

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society A*, *159*(3), 385–443.

Ingrassia, S., Minotti, S. C., & Vittadini, G. (2009). *Local statistical modeling by cluster-weighted*. ArXiv0911.2634v1.

Lilford, R., Mohammed, M. A., Spiegelhalter, D. J., & Thomson, R. (2004). Use and misuse of process and outcome data in managing performance of acute medical care: Avoiding institutional stigma. *The Lancet*, *363*, 1147–1154.

Minotti, S. C., & Vittadini, G. (2010). Local multilevel modeling for comparisons of institutional performance. In: C. Lauro, F. Palumbo, & M. Greenacre (Eds.), *Data analysis and classification: From the exploratory to the confirmatory approach* (pp. 289–295). Berlin: Springer.

Vroman Battle, M., & Rakow, E. A. (1993). Zen and the art of reporting differences in data that are not statistical significant. *IEEE Transactions on Professional Communication*, *36*(2), 75–80.

Impact Evaluation of Job Training Programs by a Latent Variable Model

Francesco Bartolucci and Fulvia Pennoni

Abstract We introduce a model for categorical panel data which is tailored to the dynamic evaluation of the impact of job training programs. The model may be seen as an extension of the dynamic logit model in which unobserved heterogeneity between subjects is taken into account by the introduction of a discrete latent variable. For the estimation of the model parameters we use an EM algorithm and we compute standard errors on the basis of the numerical derivative of the score vector of the complete data log-likelihood. The approach is illustrated through the analysis of a dataset containing the work histories of the employees of the private firms of the province of Milan between 2003 and 2005, some of whom attended job training programs supported by the European Social Fund.

1 Introduction

We develop an approach to study the effect of job training programs on the type of employment. The approach is used to analyse a longitudinal dataset containing the work histories of a large group of subjects who are resident in the Province of Milan (Italy), which includes 189 towns and municipalities.

The model we introduce may be seen as an extension of the dynamic logit model (Hsiao 2003). As such, it is based on subject-specific intercepts to account for the unobserved heterogeneity between subjects and it includes, among the regressors, the lagged response variable. This allows us to estimate the effect of the true *state dependence* (Heckman 1981), i.e., the actual effect that experiencing a certain situation in the present has on the probability of experiencing the same situation in the future. Differently from more common approaches, we assume that the random intercepts have a discrete distribution, following in this way a formulation similar to that of the latent class model (Lazarsfeld and Henry 1968). This formulation avoids to specify any parametric assumption on the distribution of the random intercepts. Among the regressors, we also include a set of dummies for having attended the training program. These dummies are time-specific; therefore, we can also evaluate whether the program has or not a constant effect during the period of observation.

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_8,

Maximum likelihood estimation of the model parameters is carried out through an Expectation–Maximization (EM) algorithm (Dempster et al. 1977). On the basis of the score vector of the complete data log-likelihood, which is obtained as a by-product of the EM algorithm, we compute the standard errors for the parameter estimates; see also Bartolucci and Farcomeni (2009).

The paper is organized as follows. In the next section we describe in more detail the dataset mentioned above. In Sect. 3 we illustrate the latent variable model and we discuss its equivalence with a model formulated following a potential outcome approach. Finally, in Sect. 4 we discuss the main results from the application of this model to the dataset described in Sect. 2.

2 The Dataset

The dataset we analyse is extracted from a database derived from the merge of two administrative archives. The first archive is made by the mandatory announcements of the employers to the public employee service registers (employment offices) operating on the Province of Milan about hiring (new contract) or firing (expired contract). It is then possible to obtain, for every employee working in a private firm, relevant data on his/her employment trajectories, such as the number of events, type and duration of the contract, sector, and qualification. Since 2000, this archive is updated at any change of the job career.

The second archive contains information about the voluntary participants to the courses supported by the European Social Fund which took place in Lombardy between 2000 and the first quarter of 2007. We selected, among the programs designed at that time, those aimed at favouring: (a) first time employment, (b) return to work, and (c) acquisition of additional skills for young employees. Most participants were aged between 18 and 35; the courses lasted on average less than six months and ranged from broadly oriented to relatively specialized topics.

With the data at hand, we chose to evaluate the impact of job-training programs on the probability of improving in the type of contractual category. We selected three main categories: (a) temporary agency, (b) temporary (fixed term), and (c) permanent (open ended) job contract. We also chose to study the impact of those programs taking place in the first quarter of 2004 and to restrict the analysis to Italian employees aged 20–35 in 2004. We then have a group of 370,869 workers: 4,146 trained subjects (1.12%) and 366,723 untrained subjects (98.88%).

Note that, from the administrative archives, the employment status of a subjects is not available if he/she is: (a) not employed, (b) employed outside the Province of Milan, (c) self-employed, or (d) employed in the public sector or with a coordinated and continued collaboration type of contract. Therefore, for each period of interest we considered a response variable having four levels: (0) if the labour state of the subject is unknown (he/she is not in the archive at this time), (1) if he/she is employed with a temporary agency contract, (2) if he/she is employed with a fixed term contract, (3) if he/she is employed with a permanent contract.

Table 1 Descriptive statistics for the covariates of trained and untrained subjects

		Trained	*Untrained*
Gender: males (%)		48.50	54.87
Age in 2003: mean		27.76	27.95
Level of Education:	missing (%)	26.75	41.87
	none or primary school (%)	0.72	1.19
	middle school (%)	21.23	23.26
	high school (%)	37.65	25.16
	college degree (%)	13.65	8.51
	higher (%)	0.00	0.02

We are interested in estimating the early effects of the training. For this reason, we consider the response variable three months before and six, nine, twelve, and fifteen months after the beginning of the program. We then have five response variables for each subject, which are denoted by $y_1, \ldots, y_5$. The categories 1–3 of such responses are ordered, with the last one corresponding to the most stable type of contract in the Italian system. We also consider the covariates listed in Table 1, where some descriptive statistics for these covariates are listed.

3 The Statistical Approach

For each subject i in the sample, $i = 1, \ldots, n$, we denote by y_{i0} and y_{i1} the labour state observed, respectively, six and three months before the first quarter of 2004 (period of the beginning of the job training program). We also denote by $y_{i2}, \ldots, y_{i5}$ the labour state observed, respectively, six, nine, twelve, and fifteen months after the first quarter of 2004.

3.1 Model Assumptions

Given the nature of the response variables, we use a model based on nested logits (Agresti 2002). For each variable we have three logits. The first one compares the probability of entering the database against not entering, i.e., category 0 against all the other categories. At nested level, we use two cumulative logits for the conditional probability of each category larger than 0, because these categories are ordered.

The model accounts for unobserved heterogeneity and state dependence by the inclusion of subject-specific intercepts and the lagged response variable among the regressors. The intercepts are treated as random parameters having a discrete distribution with k support points, which identify k latent classes in the population. The c-th class has probability denoted by π_c, $c = 1, \ldots, k$.

The model considers the first response variable y_{i0} as given, whereas the distribution of y_{i1} is modelled as follows

$$\log \frac{p(y_{i1} > 0|c_i, \mathbf{x}_{i1}, y_{i0})}{p(y_{i1} = 0|c_i, \mathbf{x}_{i1}, y_{i0})} = \alpha_{1c_i} + \mathbf{x}'_{i1}\boldsymbol{\beta}_{11} + \sum_{j=1}^{3} d_{ij0}\beta_{1,j+1},$$

$$\log \frac{p(y_{i1} > h|c_i, \mathbf{x}_{i1}, y_{i0}, y_{i1} > 0)}{p(y_{i1} \leq h|c_i, \mathbf{x}_{i1}, y_{i0}, y_{i1} > 0)} = \alpha_{2c_i} + \tau_h + \mathbf{x}'_{i1}\boldsymbol{\beta}_{21} + \sum_{j=1}^{3} d_{ij0}\beta_{2,j+1},$$

where $h = 1, 2$, $\mathbf{x}_{i1}$ is the vector of exogenous covariates at the first occasion, c_i is the latent class of subject i, and $\tau_1 \equiv 0$ to ensure identifiability. Moreover, α_{1c} and α_{2c} are the support points associated to latent class c, $c = 1, \ldots, k$, τ_2 is the shift parameter for the third logit with respect to the second, and d_{ijt} is a dummy variable equal to 1 if $y_{it} = j$ and to 0 otherwise.

For what concerns the distribution of y_{it}, $t = 2, \ldots, 5$, we assume

$$\begin{aligned} &\log \frac{p(y_{it} > 0|c_i, \mathbf{x}_{it}, y_{i,t-1}, z_i)}{p(y_{it} = 0|c_i, \mathbf{x}_{it}, y_{i,t-1}, z_i)} \\ &\quad = \alpha_{1c_i} + \mathbf{x}'_{it}\boldsymbol{\beta}_{11} + \sum_{j=1}^{3} d_{ij,t-1}\beta_{1,j+1} + z_i\gamma_{1t}, \end{aligned} \tag{1}$$

$$\begin{aligned} &\log \frac{p(y_{it} > h|c_i, \mathbf{x}_{it}, y_{i,t-1}, y_{it} > 0, z_i)}{p(y_{it} \leq h|c_i, \mathbf{x}_{it}, y_{i,t-1}, y_{it} > 0, z_i)} \\ &\quad = \alpha_{2c_i} + \tau_h + \mathbf{x}'_{it}\boldsymbol{\beta}_{21} + \sum_{j=1}^{3} d_{ij,t-1}\beta_{2,j+1} + z_i\gamma_{2t}, \end{aligned} \tag{2}$$

where $h = 1, 2$ and the vector of covariates $\mathbf{x}_{it}$ at occasion t also includes time dummies. Note that, the parameters γ_{1t} and γ_{2t}, $t = 2, \ldots, 5$, measure the effect of the job training program on each period (see the discussion below for further details).

Finally, for the binary variable z_i equal to 1 if subject i attends the job training program and to 0 otherwise, we assume

$$\log \frac{p(z_i = 1|c_i, \mathbf{x}_{i1}, y_{i0})}{p(z_i = 0|c_i, \mathbf{x}_{i1}, y_{i0})} = \alpha_{3c_i} + \mathbf{x}'_{i1}\boldsymbol{\delta}_1 + \sum_{j=1}^{3} d_{ij0}\delta_{j+1},$$

with α_{3c}, $c = 1, \ldots, k$, being support points associated to the latent classes.

In the model presented above we assume that all observable factors (represented by the covariates) and unobservable factors (represented by the random intercepts) affecting both the job status and the choice of the treatment are properly taken into account. If this assumption holds, then a causal model in the sense of Pearl (1995) results.

Indeed, causal models for observational studies similar to the present one are typically formulated following a potential outcome approach (Rubin 2005); see for instance Sianesi (2004) and Lechner and Miquel (2010). Here, the potential outcomes may be denoted by $y_{it}^{(1)}$ and $y_{it}^{(0)}$ and, for every subject i and time occasion t, indicate the type of contract if the program was or was not attended. It is worth noting that the model presented above is equivalent to a model formulated on these potential outcomes through a similar parameterisation. In a related context, the equivalence between the two formulations is derived in Bartolucci (2010) and Ten Have et al. (2003). The main assumption for this equivalence to hold is that the potential outcomes are conditionally independent of z_i given the observed covariates and the random intercepts. An important aspect is that the parameters γ_{ht} in (1) and (2) may be seen as suitable contrasts, on the logit scale, between the probabilities of certain configurations of $y_{it}^{(1)}$ and $y_{it}^{(0)}$. This enforces their interpretation as causal parameters.

3.2 Maximum Likelihood Estimation

Estimation of the model parameters is based on the maximization of the log-likelihood

$$\ell(\boldsymbol{\theta}) = \sum_i \log[p(y_{i1}, z_i, \mathbf{y}_{i2}|\mathbf{x}_{i1}, \mathbf{X}_{i2}, y_{i0})],$$

by an EM algorithm (Dempster et al. 1977). In the expression above, $\boldsymbol{\theta}$ denotes the vector of all model parameters, $\mathbf{X}_{i2} = (\mathbf{x}_{i2}, \ldots, \mathbf{x}_{i5})$, and $\mathbf{y}_{i2} = (y_{i2}, \ldots, y_{i5})'$.

As usual, this algorithm alternates two steps (E-step and M-step) until convergence and is based on the *complete data log-likelihood*. The latter may be expressed as

$$\begin{aligned}\ell^*(\boldsymbol{\theta}) &= \sum_i \sum_c u_{ic} \log[p(y_{i1}, z_i, \mathbf{y}_{i2}|c, \mathbf{x}_{i1}, \mathbf{X}_{i2}, y_{i0})\pi_c] \qquad (3)\\ &= \sum_i \sum_c u_{ic} \log[p(y_{i1}|c, \mathbf{x}_{i1}, y_{i0})] + \sum_i \sum_c u_{ic} \log[p(z_i|c, \mathbf{x}_{i1}, y_{i0})]\\ &\quad + \sum_i \sum_c u_{ic} \sum_{t>1} \log[p(y_{it}|c, \mathbf{x}_{it}, y_{i,t-1}, z_i)] + \sum_i \sum_c u_{ic} \log(\pi_c),\end{aligned}$$

where u_{ic} is a dummy variable equal to 1 if subject i belongs to latent class c and to 0 otherwise.

At the E-step, the EM algorithm computes the conditional expected value of u_{ic}, $i = 1, \ldots, n$, $c = 1, \ldots, k$, given the observed data and the current value of the parameters. This expected value is proportional to $p(y_{i1}, z_i, \mathbf{y}_{i2}|c, \mathbf{x}_{i1}, \mathbf{X}_{i2}, y_{i0})\pi_c$ and is denoted by $\hat{u}_{ic}$.

The M-step consists of maximizing the expected value of the complete data log-likelihood, obtained by substituting in (3) each u_{ic} by the corresponding expected

value computed as above. In this way we update the parameter estimates. In particular, to update the probabilities of the latent classes we have an explicit solution given by $\pi_c = \sum_i \hat{u}_{ic}/n$, $c = 1, \ldots, k$. For the other parameters we use a Fisher-scoring iterative algorithm.

A crucial point is the initialization of the EM algorithm. Different strategies may be used in order to overcome the problem of multimodality of the likelihood. As usual, it is convenient to use both deterministic and stochastic rules to choose the starting values and to take, as maximum likelihood estimate of the parameters, $\hat{\boldsymbol{\theta}}$, the solution that at convergence corresponds to the highest value of $\ell(\boldsymbol{\theta})$. In order to obtain standard errors for these estimates we rely on an approximation of the observed information matrix $\mathbf{J}(\hat{\boldsymbol{\theta}})$, which is obtained as in Bartolucci and Farcomeni (2009) on the basis of the numerical derivative of the score of $\ell(\boldsymbol{\theta})$ at $\hat{\boldsymbol{\theta}}$. This vector is directly obtained from the EM algorithm.

Finally, in order to choose the number of support points k, we use the Bayesian Information Criterion (Schwarz 1978). Therefore, we compute the index $BIC = -2\ell(\hat{\boldsymbol{\theta}}) + g \log(n)$, where g is the number of non-redundant parameters, for increasing values of k until the value of BIC slightly decreases or increases with respect to the one previously computed. The selected value of k is the one corresponding to the last fitted model (in the first case) or to the previous one (in the second case), so that this model has the smallest BIC among all the fitted models.

4 Results

We applied the proposed approach to the dataset described in Sect. 2. According to the criterion described above, we selected the model with $k = 4$ latent classes. This model has 53 parameters, maximum log-likelihood equal to $-1{,}027{,}004$, and BIC equal to 2,054,688. This last value is much lower than that of the model without unobserved heterogeneity which has maximum log-likelihood equal to $-1{,}043{,}618{,}41$ with parameters, and BIC equal to 2,087,762. For both models, the parameter estimates are reported in Tables 1 and 2. Note that, for the model with unobserved heterogeneity, the four classes have estimated probabilities equal to 0.036, 0.090, 0.860, and 0.014.

The most interesting aspect is that the estimates of the parameters γ_{ht}, which measure the dynamic impact of the training program, considerably change when unobserved heterogeneity is taken into account, i.e., when we use four latent classes instead of one. In particular, the estimates for the first logit ($h = 1$), which concerns the probability of entering the archive, are always negative with $k = 1$ and become positive with $k = 4$. Less evident is the difference in the estimates of these parameters for the second and third logits ($h = 2$). Under both models, these estimates indicate that the training program has a significant effect on the probability of improving in the contractual level only for the first period after the beginning of the program ($t = 2$). There is no evidence of a significant effect for the other periods.

Table 2 Estimates of the parameters for the conditional distribution of the response variables given the latent variable

Effect		*First logit*							
		$k = 4$				$k = 1$			
		Estimate	s.e.	t-statistic	p-value	Estimate	s.e.	t-statistic	p-value
intercept[a]	$(\bar{\alpha}_1)$	−1.127	–	–	–	−1.222	–	–	–
time dummies	(β_{111})	0.402	0.007	59.27	0.000	0.414	0.007	61.01	0.000
	(β_{112})	0.490	0.008	64.67	0.000	0.427	0.007	60.56	0.000
	(β_{113})	0.458	0.008	56.25	0.000	0.362	0.007	50.30	0.000
	(β_{114})	−0.068	0.008	−8.46	0.000	−0.081	0.007	−11.36	0.000
gender[b]	(β_{115})	−0.020	0.006	−3.63	0.001	−0.025	0.005	−5.54	0.000
age[c]	(β_{116})	0.029	0.001	44.72	0.000	0.024	0.001	45.22	0.000
dummy educ.[d]	(β_{117})	0.137	0.014	9.59	0.000	0.186	0.012	16.00	0.000
education	(β_{118})	0.054	0.005	11.20	0.000	0.075	0.004	19.04	0.000
lag response	(β_{12})	2.213	0.012	190.89	0.000	2.186	0.010	217.87	0.000
	(β_{13})	2.521	0.008	330.06	0.000	2.642	0.007	394.91	0.000
	(β_{14})	3.857	0.007	564.71	0.000	3.818	0.006	673.75	0.000
training	(γ_{12})	1.132	0.067	16.89	0.000	−0.264	0.041	−6.49	0.000
	(γ_{13})	0.635	0.072	8.81	0.000	−0.919	0.044	−20.84	0.000
	(γ_{14})	0.758	0.071	10.60	0.000	−0.819	0.044	−18.49	0.000
	(γ_{15})	1.371	0.070	19.63	0.000	−0.339	0.044	−7.63	0.000
		Second, third logits							
intercept[a]	$(\bar{\alpha}_2)$	4.693	–	–	–	3.421	–	–	–
shift	(τ_2)	−4.398	0.010	−441.84	0.000	−3.647	0.007	−518.28	0.000
time dummies	(β_{211})	0.474	0.011	43.77	0.000	0.480	0.010	48.80	0.000
	(β_{212})	−0.100	0.011	−8.86	0.000	−0.128	0.010	−12.63	0.000
	(β_{213})	0.118	0.011	10.37	0.000	0.072	0.010	7.11	0.000
	(β_{214})	0.004	0.012	0.30	0.023	−0.028	0.011	−2.60	0.009
gender[b]	(β_{215})	0.117	0.007	16.42	0.000	0.094	0.006	14.78	0.000
age[c]	(β_{216})	0.019	0.001	21.89	0.000	0.019	0.001	24.48	0.000
dummy educ.[d]	(β_{217})	0.212	0.018	11.66	0.000	0.235	0.016	14.59	0.000
education	(β_{218})	−0.013	0.006	−2.02	0.060	−0.043	0.005	−7.75	0.000
lag response	(β_{22})	−5.760	0.017	−331.28	0.000	−5.056	0.014	−348.97	0.000
	(β_{23})	−2.090	0.009	−226.30	0.000	−1.493	0.008	−194.60	0.000
	(β_{24})	4.455	0.013	348.73	0.000	4.683	0.012	396.20	0.000
training	(γ_{22})	0.216	0.069	3.12	0.055	0.183	0.062	2.96	0.003
	(γ_{23})	0.202	0.093	2.17	0.258	0.138	0.083	1.67	0.094
	(γ_{24})	0.099	0.094	1.05	0.978	−0.002	0.082	−0.03	0.979
	(γ_{25})	0.102	0.096	1.06	0.975	−0.062	0.082	−0.75	0.454

[a]average of the estimated intercepts.
[b]dummy equal to 1 for a male and 0 for a female.
[c]minus average age.
[d]dummy for the category of education missing.

For what concerns the parameters measuring the effect of the individual covariates on the response variables, we do not observe a great difference between the model with four latent classes and that with one latent class. For both models, we note that, males tend to improve more easily than females in the contractual level.

Table 3 Estimates of the parameters for the conditional probability of attending the training program given the latent variable

Effect		$k = 4$				$k = 1$			
		Estimate	s.e.	t-statistic	p-value	Estimate	s.e.	t-statistic	p-value
intercept[a]	($\bar{\alpha}_3$)	−5.098	–	–	–	−4.458	–	–	–
gender[b]	(δ_{11})	−0.161	0.034	−4.76	0.000	−0.153	0.032	−4.84	0.000
age[c]	(δ_{12})	−0.003	0.004	−0.73	0.027	−0.001	0.004	−0.31	0.759
dummy educ.[d]	(δ_{13})	0.114	0.084	1.35	0.910	0.031	0.079	0.39	0.694
education	(δ_{14})	0.316	0.027	11.84	0.000	0.249	0.025	10.04	0.000
init. period	(δ_2)	−1.152	0.111	−10.42	0.000	−0.678	0.104	−6.51	0.000
	(δ_3)	−1.434	0.069	−20.72	0.000	−0.921	0.064	−14.50	0.000
	(δ_4)	−1.304	0.043	−30.58	0.000	−0.813	0.035	−23.28	0.000

[a]average of the estimated intercepts.
[b]dummy equal to 1 for a male and 0 for a female.
[c]minus average age.
[d]dummy for the category of education missing.

Moreover, age has a positive effect on the first logit and also on the second and third logits, whereas the number of years of education has a significant positive effect only on the first logit. A strong state dependence is also observed since all the parameters associated to the lagged responses are highly significant, indicating a strong persistence on the same type of contract.

Finally, from the results in Table 3, it emerges that the covariates that have a significant effect on the propensity to attend the job training program are gender, years of education, and the response at the initial period. In particular, females have a higher propensity to attend the program, as well as subjects with higher educational level and with a less favourable contract position at the beginning of the period of observation.

Acknowledgements We acknowledge the financial support from the 'Einaudi Institute for Economics and Finance' (Rome - IT) and from the PRIN 2007 grant. We are grateful to Prof. Mario Mezzanzanica and Dr. Matteo Fontana, CRISP, University of Milano-Bicocca, for providing the dataset.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, New Jersey: Wiley.

Bartolucci, F. (2010). On the conditional logistic estimator in two-arm experimental studies with non-compliance and before–after binary outcomes. *Statistics in Medicine, 29*, 1411–1429.

Bartolucci, F., & Farcomeni, A. (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *Journal of the American Statistical Association*, *104*, 816–831.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society: Series B*, *39*, 1–38.

Heckman, J.J. (1981). Heterogeneity and state dependence. In S. Rosen (Ed.), *Studies in labor markets* (pp. 91–139). Chicago: University of Chicago Press.

Hsiao, C. (2003). *Analysis of panel data* (2nd ed.). New York: Cambridge University Press.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Lechner, M., & Miquel, R. (2010). Identification of the effects of dynamic treatments by sequential conditional independence assumptions. *Empirical Economics, 39*, 111–137.

Pearl, J. (1995). Causal diagrams for empirical research (with discussion). *Biometrika*, *82*, 669–710.

Rubin, D. B. (2005). Causal inference using potential outcomes: design, modeling, decisions. *Journal of the American Statistical Association*, *100*, 322–331.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Sianesi, B. (2004). An evaluation of the Swedish system of active labour market programs in the 1990s. *Review of Economics and Statistics*, *86*, 133–155.

Ten Have, T. R., Joffe, M., & Cary, M. (2003). Causal logistic models for non-compliance under randomized treatment with univariate binary response. *Statistics in Medicine*, *22*, 1255–1283.

Part II
Data Analysis in Economics

Analysis of Collaborative Patterns in Innovative Networks

Alfredo Del Monte, Maria Rosaria D'Esposito, Giuseppe Giordano, and Maria Prosperina Vitale

Abstract This paper focuses on territorial innovative networks, where a variety of actors (firms, institutions and research centers) are involved in research activities, and aims to set up a strategy for the analysis of such networks. The strategy is twofold and relies, on the one hand, on the secondary data available from administrative databases and, on the other, on survey data related to the organizations involved in innovative networks. In order to describe the peculiar structures of innovative networks, the proposed strategy adopts the techniques suggested in the framework of Social Network Analysis. In particular, the main goal of the analysis is to highlight the network characteristics (interactions between industry, university and local government) that can influence network efficiency in terms of knowledge exchange and diffusion of innovation. Our strategy will be discussed in the framework of an Italian technological district, i.e., a type of innovative network.

1 Introduction

During the few last decades, regional systems for innovation based on extensive interactions between industry, university and local government have shown high rates of growth (Etzkowitz and Leydesdorff 2000). Different kinds of interaction of these three subsystems have produced knowledge generation and innovative diffusion in developed and under developed areas and many papers have shown that the level of individual R&D is declining in the level of collaborative activity and that, in many cases, collaboration leads to efficient networks (among the others Goyal and Moraga-Gonzalez (2001), Jackson (2008)). Hence institutional policies have been implemented in many industrial economies to set up innovative networks (Goyal and Moraga-Gonzalez 2001, Giuliani et al. 2005, Cantner and Graf 2006).[1] However,

[1] Innovative networks can be conceived as: (i) *networks of institutions*, whose interactions determine the innovative performance of domestic firms or more specifically as a set of distinct institutions which jointly and individually contribute to the development and diffusion of technologies and which provide the framework within which governments generate and implement

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_9,

very little attention has been paid by the economic literature and the evaluation agencies to measuring the performance of such policies.

To this aim, two factors need to be considered. Networks may differ considerably in their structure (complete, star, partially connected network, circular network, empty network, see for instance (Wasserman and Faust 1994, Jackson 2008) and in the type of links that connect two or more firms. Furthermore, collaboration agreements and alliances may vary greatly in form, ranging from loose and informal agreements (such as a "memorandum of understanding") all the way to the formation of common legal organizations with strong equity ties (such as an RJV). The nature of the activities also varies. Collaboration might involve the sharing of production facilities, the sharing of information concerning new technologies and new products or the production of specialized components.

Furthermore, policies aiming to support the creation of network structures might have different welfare effects in relation to the characteristics of the environment. In less developed areas, where the level of social trust is weak and small firms are not used to collaborating, networks of large firms (often not local) and universities could be more effective than networks of small firms. On the other hand, when the level of social capital is more developed, networks of small firms and local authorities could be associated with strong economic performance (Putnam 1995, Knack and Keefer 2003). It is therefore important to develop a methodology that makes it possible to understand what kind of network is more likely to have a positive impact in a given region and how to measure the performance of implemented policies.

Given that the techniques proposed in the framework of Social Network Analysis (Wasserman and Faust 1994) take into account information exchange and flows of knowledge among different organizations, it may be useful to adopt them both to evaluate the performance of actions implemented by technological districts and to describe the impact of public policies for territorial innovation and knowledge diffusion. In this paper we study a particular type of innovative network: the technological district. We set up a methodology to analyze the policy which started in 2002 in order to support the start-up of technological districts in Italy.[2] This methodology will be applied to a network of firms and public and private research centers located in the Campania Region of Southern Italy.

The paper is organized as follows: in Sect. 2 a brief review of the literature on economic incentives that could be used to increase the information exchange and flows of knowledge among firms is presented; in Sect. 3 the strategy proposed for innovative network analysis in terms of *complete networks* to analyze secondary data structure and *ego-centered networks* for survey data is briefly presented. Section 4 contains the study of the collaboration network in the chosen technological district.

---

policies to influence the innovation process (Giuliani et al. 2005); (ii) *networks of actors*, which cooperatively engage in the creation of new ideas and then economize on the results (Cantner and Graf 2006); (iii) *network of firms* with a high level of inter firm collaboration in R&D activity (Goyal and Moraga-Gonzalez 2001).

[2] The two most important laws outlining the procedures within each Italian region that could implement a technological district are L.317/91 and L.140/99.

2 Economic Incentives to Exchange Information and Knowledge among Firms

The advantages of cooperation between firms derives from: (i) the strong complementarity of the R&D activities; (ii) the feasible coordination among the firms. These factors bring about a reduction in the costs of R&D activities and avoid the duplication of many fixed costs. Moreover, there are information benefits that depend on the specific structure adopted by the firms. These advantages afford the possibility to increase the diffusion of information, monitor competitors and reduce the risk of free-riding. Furthermore, when faced with technological spillover, only the firms that jointly invest in R&D and RJV internalize these benefits, while those that under-invest in R&D suffer a negative impact on their performance and welfare (Kamien et al. 1992). The cost of customizing inputs and the probability of strategic opportunistic behaviour on the part of the partners decrease when many actors are directly connected to each other.

The dimension of the network is a factor often considered as relevant (Katz 1986). In general, we can say that cooperative agreements among firms in technologically advanced sectors, such as pharmaceuticals, are often very efficient (Orsenigo et al. 2001). Nevertheless, following the coalition games approach, when the spillovers affect out-of-the-network firms, Yi and Shin (2000) prove that a free-riding behavior can still be optimal for the firm. Public enforcement is necessary in order to allow innovative networks to be sustainable and stable.

Other theoretical studies focus on the cooperation between firms and public and/or private institutes of research, such as universities (Aghion et al. 2005, Lacetera 2009). The production of knowledge is seen as the result of specific efforts by the actors (firms, universities, local authorities) that have to decide if it is better to coordinate their efforts or act strategically.

Furthermore, many papers (Teece 1981, Audretsch and Feldman 1996, Breschi and Lissoni 2001) outline the importance, due to spatial externalities, of geographical proximity in the exchange and knowledge flows among actors. The cost of transferring information and knowledge is in fact greatly reduced when actors belong to a localized network.

3 A Social Network Analysis Approach to Describe the Collaborative Patterns in Innovative Networks

By taking into account information exchange and flows of knowledge among different organizations, the techniques proposed in the framework of Social Network Analysis (SNA) can be fruitfully employed to: (i) describe and analyze the collaborative patterns in innovative networks; (ii) define the role and position of organizations in the net according to actor-level measures; (iii) identify

groups of homogeneous organizations; (iv) observe collaboration in the network in evolutionary terms, highlighting various net configurations.

The strategy proposed in the following is based on two kinds of information. On the one hand, we consider secondary data available from administrative databases to define a complete network. On the other, we rely on survey data able to define ego-centered networks (e.g., Hanneman and Riddle 2005).

In order to define a complete network that can describe the collaboration among organizations, our strategy requires the definition of network boundaries by identifying the actors and the type of relationships. The actors are firms and institutions involved in the activities of innovative networks, while the links (relational data) can be defined according to their joint participation in research projects. An *affiliation matrix* $\mathbf{F}_{(n\times p)}$, where n and p are respectively the number of organizations and research projects, and a symmetric valued *adjacency matrix* $\mathbf{A}_{(n\times n)}$, that describes the collaboration network among organizations according to the participation in research projects, can be defined (Wasserman and Faust 1994). From $\mathbf{A}$, some network indices of interest (such as *density, centrality*, etc.) and actor-level measures can be computed. To define homogeneous groups of actors that have the same (or similar) pattern of ties, we look for a complete clustering by means of Hierarchical Cluster Analysis carried out on a dissimilarity matrix $\mathbf{D}_{(n\times n)}$, obtained by computing a suitable dissimilarity measure (Sneath and Sokal 1973) between pairs of actors. Furthermore, clusters of actors that share within and between common patterns of ties are determined by the Blockmodeling techniques (e.g., Lorrain and White 1971, Doreian et al. 2005). In particular, it is possible to assemble a set of blocks into a blockmodel to discern the real structure present in the data, such as cohesive, core-periphery or hierarchical configurations.

Finally, in our strategy we define ego-centered networks in order to observe how individual actors are embedded in local structures. The relational data are collected by survey data through the administration of a questionnaire to organizations involved in innovative networks. The evolution of network configurations can be appraised through the recognition of existing links among organizations before, during and after the establishment of an innovative network.

In the following the proposed method will be applied to study the collaboration network among private and public actors involved in a technological district whose projects have been financially supported during the years 2005–2007. Alternative applications of the proposed method on different kinds of networks with a larger number of actors and a longer period of analysis could be envisaged.

4 The Collaboration Network in an Italian Technological District

In this section we discuss the main results from the analysis of the collaboration network in a technological district in the Campania Region of Southern Italy. This district is composed of 24 organizations (firms, university departments and public

research institutes) and was established in February 2004. Private actors are large firms, some of which already have research units in Campania.

4.1 The Study of Collaborative Patterns in the Technological District: The Complete Network

Starting from available data, we first define a collaboration network among the 24 organizations. The links are defined according to their joint participation in research projects, financially supported by public funding in the years 2005–2007.

According to the participation in research projects,[3] we define a complete network to describe the collaboration among organizations (Firms; public research institutes (R_Inst) and university departments (U_Dep)) involved in the technological district. The network visualization (Fig. 1a) shows the presence of a group of organizations (R_Inst1, U_Dep4, Firm1, Firm2 and Firm6) which participate in a large number of projects and some organizations (mainly university departments) which rarely cooperate with the other organizations. The network density, equal to 0.61, reveals a quite strong cohesion in the net. The analysis of actor-level measures, based on centrality indices[4] highlights the presence of two public research centers that have an important position in the net (U_Dep4 and R_Inst1). Nevertheless, there are some private firms (e.g., Firm1, Firm2 and Firm6) that play a central role with a relevant number of relationships. The identification of homogeneous groups is carried out by looking at a complete clustering over the set of organizations by means of Hierarchical Cluster Analysis, conducted on dissimilarity matrix $\mathbf{D}_{(24\times24)}$. We derived a partition into three groups which are further analyzed by means of Blockmodeling techniques for valued adjacency matrix (Ziberna 2007).[5] Figure 1b highlights the presence of a core-periphery configuration in the collaboration network: one dense, small cohesive group of public and private leaders (U_Dep4, R_Inst1, Firm1, Firm2 and Firm6) that maintain strong relationships within and between clusters; one semi-peripheral group composed of private firms that collaborate with the leaders and have few ties with the other group and finally one sparse, unconnected peripheral group, consisting mainly of university departments, characterized by low participation in the research projects.

[3] Starting from a binary *affiliation matrix* $\mathbf{F}_{(24\times12)}$, with 24 organizations and 12 research projects, we define a valued *adjacency matrix* $\mathbf{A}_{(24\times24)}$. In matrix $\mathbf{A}$ the links are defined according to their joint participation in research projects of pairs of organizations.

[4] To compute the centrality indices we consider a 0/1 version of the *adjacency matrix* $\mathbf{A}$, where the threshold value to define the presence of a link is equal to 1.

[5] Euclidian distance, Ward agglomerative method and Blockmodeling analysis are performed on the valued *adjacency matrix* $\mathbf{A}$ using the package Blockmodeling of R software implemented by Ziberna Ziberna (2007). In particular, the "Val" approach is considered with a threshold value m equal to 5.

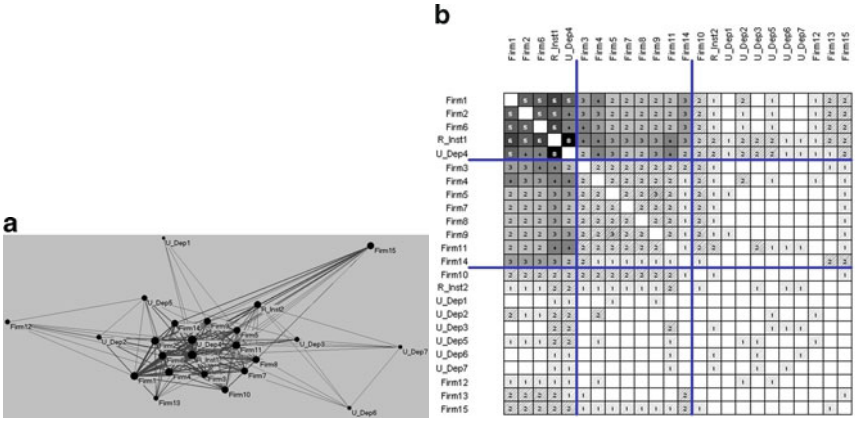


Fig. 1 (**a**) Collaboration network among the 24 organizations, according to the participation in research projects. (**b**) Permuted adjacency matrix A obtained according to the three groups of equivalent organizations

4.2 A Survey on the Organizations of the Technological District: The Ego-Network Approach

To understand some peculiarities in the collaboration behavior of the 24 organizations involved in the technological district, we describe how individual actors are embedded in local structures. Hence, we define *ego-centered networks*, with an individual focal organization (*ego*) and all organizations with whom ego has connections (*alters*). A survey[6] has been conducted in order to define the *ego-centered networks*, where the actors are the 24 organizations and the ties describe the participation in common research projects according to three time occasions before, during and after the establishment of the technological district (*ex-ante*, *actual* and *ex-post* networks). In particular, we highlight the structure of some peculiar *egos*. The firms that have a central position in the network are Firm1, Firm2, Firm4 and Firm6. They represent the strongest organizations in terms of *size* and *density*. Therefore they play a fundamental role in the collaboration network to communicate with each other (*betweenness*) and to mediate the communication flow (*brokerage*). The analysis of ego-networks highlights three typologies that can be considered typical structures in the domain of all 24 organizations. The first typical pattern, as exemplified by Firm1 (Fig. 2) and Firm6, is characterized by an increasing number of relationships in the three network configurations (*ex-ante*, *actual* and *ex-post* networks). The second one, as exemplified by Firm2 (Fig. 2) and Firm4, shows relative stability in

[6] The data, gathered by means of a survey conducted from September 2008 to February 2009 through a questionnaire administered to organizations, are related to production processes, training activities and participation in research projects before, during and after the establishment of the technological district.

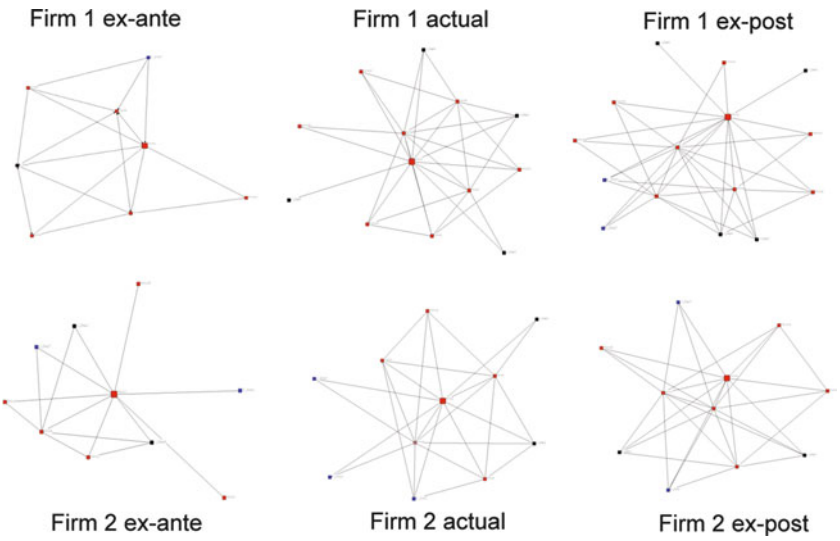


Fig. 2 A comparison of two ego-networks in the three time occasions: Firm1 shows an increase in the number of degrees and density in the three time occasions. Firm2 maintains a relatively stable number of degrees before, during and after the establishment of technological district

the links evolution. Finally, the third structure refers to some firms that show poor collaborative behaviors in the technological district.

5 Some Final Remarks

The main results obtained by analyzing the complete network and ego-networks defined for the actors of the technological district seem promising to assess the collaborative behaviors activated by private firms and public research institutions for innovation diffusion and knowledge exchange.

Possible alternative applications of the proposed method, on different kinds of networks and on wider networks are possible, given the characteristics of the methodology, which is not bounded to a number of actors in a given interval of values or to a particular typology of the network. The analysis could be carried out over a longer period of time than the one considered in the present paper by appropriately defining what is ex-ante and what is ex-post. For instance the proposed methodology could be useful in analysing innovative networks of local organizations and to define indicators of network based policies (including performance assessment of the technological districts financially supported by the Ministry of Education, University and Research in Italy).

References

Aghion, P., Dewatripont, M., & Stein, J. C. (2005). *Academic freedom, private-sector focus, and the process of innovation*. Harvard Institute of Economic Research, Discussion Paper No. 2089. Available via DIALOG. http://ssrn.com/abstract=775064.

Audretsch, D. B., & Feldman, M. P. (1996). R&D spillovers and the geography of innovation and production. *American Economic Review*, *86*, 630–640.

Breschi, S., & Lissoni, F. (2001). Knowledge spillovers and local innovation sistems: A critical survey. *Industrial and Corporate Change*, *10*, 975–1005.

Cantner, U., & Graf, H. (2006). The network of innovators in Jena: An application of social network analysis. *Research Policy*, *35*, 463–480.

Doreian, P., Batagelj, V., & Ferligoj, A. (2005). *Generalized blockmodeling*. Cambridge: Cambridge University Press.

Etzkowitz, H., & Leydesdorff, L. (2000). The dynamics of innovation: From National Systems and "Mode 2" to a Triple Helix of university-industry-government relations. *Research Policy*, *29*, 109–123.

Giuliani, E., Rabellotti, R., & van Dijk, M. P. (2005). *Clusters facing competition: The importance of external linkages*. Hampshire, UK: Aldershot Ashgate.

Goyal, S., & Moraga-Gonzalez, J. L. (2001). R&D Networks. *RAND Journal of Economics*, *32*, 686–707.

Hanneman, R. A., & Riddle, M. (2005). *Introduction to social network methods*. University of California, Riverside. Available via DIALOG. http://faculty.ucr.edu/~hanneman/

Jackson, M. O. (2008). *Social and economic networks*. Princeton: Princeton University Press.

Kamien, M. I., Muller, E., & Zang, I. (1992). Research joint ventures and R&D cartels. *The American Economic Review*, *82*, 1293–1306.

Katz, M. (1986). An analysis of cooperative research and development. *Rand Journal of Economics*, *17*, 527–543.

Knack, S., & Keefer, P. (2003) Does social capital have an economic payoff? A cross-country investigation. In S. Knack (Ed.), *Democracy, governance and growth* (pp. 252–288). Ann Arbor: The University Michigan Press.

Lacetera, N. (2009) *Different missions and commitment power in R&D organizations: Theory and evidence on Industry-University alliances. Journal of Organization science*, *30*(3), 565–582. Available via DIALOG. http://ssrn.com/abstract=1085924

Lorrain, F., & White, H. C. (1971). Structural equivalence of individuals in social networks. *Journal of Mathematical Sociology*, *1*, 49–80.

Orsenigo, L., Pammolli, F., & Riccaboni, M. (2001). Technological change and network dynamics: Lessons from the pharmaceutical industry. *Research Policy*, *30*, 485–508.

Putnam, R. D. (1995). Bowling alone: America's declining social capital. *Journal of Democracy*, *6*, 64–78.

Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy: The principles and practice of numerical classification*. San Francisco: Freeman.

Teece, D. J. (1981). The market for know-how and the efficient international transfer of technology. *Annals of the American Academy of Political and Social Science*, *458*, 81–96.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge: Cambridge University Press.

Yi, S., & Shin, H. (2000). Endogenous formation of research coalitions with spillovers. *International Journal of Industrial Organization*, *18*, 229–256.

Ziberna, A. (2007). Generalized blockmodeling of valued networks. *Social Networks*, *29*, 105–126.

The Measure of Economic Re-Evaluation: a Coefficient Based on Conjoint Analysis

Paolo Mariani, Mauro Mussini, and Emma Zavarrone

Abstract During the last 40 years conjoint analysis has been used to solve a wide variety of concerns in market research. Recently, a number of studies have begun to use conjoint analysis for the economic valuation of non-market goods. This paper discusses how to extend the conjoint analysis area of application by introducing a coefficient to measure economic re-evaluation on the basis of utility scores and the relative importances of attributes provided by conjoint analysis. We utilise the suggested coefficient for the economic valuation of a typical non-market good, such as a worldwide cultural event, to reveal the trade offs between its attributes in terms of revenue variation. Our findings indicate the most valuable change to be made to the existing status quo to generate economic surplus.

1 Introduction

Conjoint analysis is a standard tool for studying customer/user preferences (Cattin and Wittink 1982), producing an estimation of attribute utilities and a definition of their relative importance, thus, providing a total utility of the product status quo. We can, however, hypothesise a change in this status quo to evaluate such a change in terms of total utility variation. In so doing, an issue related to economic valuation of the introduced change arises. Valuation of this kind can be particularly useful when concerned with non-market goods which do not have a well-defined purchasing process given their characteristical lack of rivalry and exclusiveness (Sanz et al. 2003). In this paper, we propose a new coefficient to evaluate the revenue after a change in the combination of attributes of a given non-market good. This coefficient works both on the utility score and on the relative importance value assigned to the various attributes by using the OLS multiple regression as an estimation procedure. In our application of the coefficient of re-evaluation we refer to the part-worth function model as it represents the most general preference model, however, the use of an alternative preference model is not excluded although it is not discussed here. The remainder of the paper is organised as follows: Sect. 2 outlines the framework of conjoint analysis use for non-market economic valuation; Sect. 3 introduces

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_10,

the conjoint based coefficient of economic re-evaluation; in Sect. 4, we apply the methodology defined in the previous section to a ranking conjoint survey carried out from July to November 2007 on visitors of "Venice and Islam 828-1797", a cultural event that was located in Venice; Sect. 5 is devoted to final remarks.

2 Using Conjoint Analysis for Non-market Economic Valuation

Conjoint analysis is a widely used technique for investigating consumer choice behaviour in commercial studies (Cattin and Wittink 1982). Since its introduction, conjoint analysis has become a popular tool for decomposing the structure of consumer preference by estimating the parameters of a preference model according to the consumer evaluation of a set of alternatives which differ in terms of levels of the attributes that specify the composition of those alternatives available in the set (Green and Srinivasan 1978). In marketing research, the task of conjoint analysis is directed towards revealing the trade-offs that consumers are willing to make among attributes and the prediction of how those consumers will react to the introduction of new products on the market. Recently, conjoint analysis has been developed as a state-preference method for the economic valuation of changes in non-market goods, such as environmental commodities or cultural services (Roe et al. 1996, Willis and Snowball 2009). As price is commonly included as an attribute, conjoint analysis provides an estimate of the relative importance of price with respect to the remaining attributes. Since any stated preference technique serves the scope of estimating non-market values, the result of conjoint analysis provides information concerning consumer willingness to pay for any change in the level of an attribute. However, conjoint analysis is only one of the methods for estimating willingness to pay for non-market goods. With the aim of investigating willingness to pay for environmental or cultural goods, contingent valuation is often used as a stated preference elicitation technique (Desvousges et al. 1993, Sanz et al. 2003). Contingent valuation is a method by which the respondent is asked about his maximum willingness to pay for a change in a non-market good. Thus, the process of making decisions differs from that required by the conjoint format in which the respondent is asked to compare alternatives which have a pre-specified price (Irwin et al. 1993, Stevens et al. 2000). As noted in Boxall et al. (1996), if respondents tend to ignore alternatives when making decisions about price in accordance with the contingent valuation method, willingness to pay resulting from the contingent valuation method may be quite different from those obtained by using conjoint methodology. Although a variety of conjoint models have been developed in literature (Green and Srinivasan 1990), most conjoint based non-market valuations adopt a choice modelling approach concerned with the random utility theory (Mazzanti 2003). The random utility model often works on the probability of choosing the most preferred choice from among a set of known alternatives. However, this model does not fully exploit all of the information contained in the conjoint ranking format. When respondents

are asked to express the exact rank order of choice set alternatives, the additional information about ordinal ranking of the remaining alternatives beyond the first choice is not utilised. However, a typical request of conjoint surveys is either to rank alternatives or to rate alternatives on an integer scale (Chapman and Staelin 1982, Mackenzie 1993). Standard conjoint analysis uses OLS to exploit this additional information thereby regressing individual responses to a linear function of the attributes.

3 A Conjoint Based Model for Economic Re-Evaluation

The application of conjoint analysis requires several steps and various options are available for each of these steps. The first step is the choice of the preference model that relates to the respondent preferences and to the levels of the attributes of the available alternatives. Three main preference models are used in conjoint studies: the part-worth function model (piecewise linear), the vector model (linear), and the ideal point model (linear plus quadratic). We consider the part-worth function model as it is the most general model providing a specific utility value for each level of the considered attributes usually referred to as part-worth utility (Green and Srinivasan 1990). Having defined the preference model, we go on to choose the data collection method and the selection of alternatives to estimate the part-worth utilities. As the detailed discussion of such phases exceeds the objectives of the present argumentation, we presume that both the data and the alternatives have already been defined so we then proceed to estimate the part-worth utilities by using the multiple linear regression in accordance with the metric conjoint analysis approach. This estimation procedure can be adopted either in presence of rank order preferences or when a rating scale is used as a response mode (Carmone et al. 1978). Obviously, the higher the utility value is, the more the corresponding attribute level is appreciated. Otherwise, the lowest utilities are associated with the least preferred levels. Therefore, we can estimate the total utility of any hypothetical combination of levels by summing the corresponding utility values. According to this approach, the respondent overall evaluation of the k^{th} alternative in terms of utility is:

$$p_k = \beta_0 + \sum_{j=1}^{J} \sum_{h=1}^{H_j} \beta_{jh} a_{jh} \tag{1}$$

where β_{jh} is the part-worth utility associated with the dichotomous variable a_{jh} denoting the presence ($a_{jh} = 1$) or the absence ($a_{jh} = 0$) of the h^{th} level of the j^{th} attribute, and β_0 is a constant.

In this framework, a coefficient of economic re-evaluation for a hypothetical change occurring in the combination of the attribute levels can be defined by a pairwise comparison of the total utility values of two alternatives which differ by that changed attribute level. Being p_b the sum of the part-worth utilities associated with the status quo of the good, p_j denotes the sum of the utility scores related to a

specific combination of the attributes with the j^{th} attribute modification (alternative j). Thus, we can calculate the total utility variation linked to a modification of the attribute j with respect to the status quo level of that attribute. M_j indicates the ratio which results by dividing the difference between the total utility of the alternative j and the status quo one by the total utility assigned to the status quo, formally

$$M_j = \frac{p_j - p_b}{p_b} \tag{2}$$

where p_b is assumed to be diverse from 0. The result of this ratio indicates whether the status quo modification will generate a loss or a gain in terms of total utility. It is evident that a zero value for M_j represents the indifferent situation between loss and gain in terms of total utility. However, the utility modification arising from an attribute level modification can be considered more or less important by consumers. Consequently, such attribute level modification can have a more important economic impact than a utility modification which has a similar intensity but which involves a less important attribute. As a solution, we use the relative importance of the modified attribute as an indicator of the attribute impact on the overall utility determination (Srinivasan and Shocker 1973). The range of the utility values (highest to lowest) for each attribute provides an indicator of how important the attribute is with respect to the remaining attributes. Attributes with larger utility ranges play a more important role than those with smaller ranges. For each respondent i, the relative importance of each attribute can be computed by dividing its utility range by the sum of all utility ranges in order to obtain a ratio-scaled importance measure as follows:

$$I_{ij} = \frac{\max\left(u_{ij}\right) - \min\left(u_{ij}\right)}{\sum_{j=1}^{J}\left[\max\left(u_{ij}\right) - \min\left(u_{ij}\right)\right]} \tag{3}$$

where J is the number of the considered attributes and u_{ij} is the set of part-worth utilities referred to various levels of attribute j. We compute the importance values separately for each respondent and then average them rather than calculating the importance values from the summary part-worth utilities obtained by averaging the part-worth utilities for all the respondents. This procedure permits to summarize the attribute importance values by considering how the attribute importance values vary over the various respondent preference structures. Therefore, the average importance values generally differ from those obtained by using the summary part-worth utilities. We can express these relative importance values in terms of decimal fractions whose sum is one. Thus, we can introduce the relative importance of the modified attribute in (2), so that the coefficient formulation becomes

$$MI_j = M_j * I_j \tag{4}$$

where I_j is the relative importance assigned to attribute j. If we admit that p_b can be negative, we then use a minus before MI_j and the general formulation of the coefficient becomes

$$MI_j = \begin{cases} \frac{p_j - p_b}{p_b} * imp_j & \text{if } p_b > 0 \\ \\ \frac{p_b - p_j}{p_b} * imp_j & \text{if } p_b < 0 \end{cases} \tag{5}$$

Expression (5) provides an overall utility variation measure for one change in the status quo profile of a given multi-attribute good. Due to its relative importance based weighting system, the coefficient MI_j enables the valuation of the hypothetical changes occurring in the status quo profile one at a time.

If we refer to a typical public good, such as a cultural exhibit or a theatrical performance, we can use formula (5) to estimate the variation of the total revenue generated by assuming a change in the status quo profile of that cultural activity. Having defined the total revenue associated with the status quo profile as π, we can introduce this monetary amount into formula (5) to obtain a valuation of the revenue variation for a change in the attribute combination of the public good. In so doing, V_j denotes the amount of the revenue variation, as is shown by the following equation

$$V_j = MI_j * \pi \tag{6}$$

The revenue variation in (6) is estimated by supposing that the monetary factor of a cultural good (price or admission charge) varies in proportion to the change of the total utility of that good. We argue that if the price of a cultural event reflects on how an user evaluates the combination of attributes of the cultural event in terms of utility, the economic value of a change in the combination of attributes can be expressed as a function of the utility and of the importance of the modified attribute. In addition, we notice that conjoint analysis serves the scope of approximating the real structure of consumer preference as only partial knowledge of consumer preference can be known. We, therefore, suggest the use of the coefficient of economic re-evaluation as a monetary indicator which approximates the impact of a given utility change in monetary terms.

4 Case Study

We refer to a survey concerning the cultural event "Venice and Islam 828-1797" to test the coefficient of re-evaluation. After Paris and New York, this large-scale exhibition on the relationship between Venice and the world of Islam was hosted in Venice itself in the symbolic Doge's Palace (28 July–25 November 2007).

4.1 Survey Design and Data Collection Method

The sample comprises 501 respondents who were interviewed after the visit. Data was collected by face to face interviews in which each respondent was asked to

rank eight alternative profiles presented in the form of a questionnaire. These alternatives were carried out from a full factorial design produced by a permutation of all attribute levels (Placket 1946). Each alternative is described by four attributes: admission charge, location, modality of gathering information about the exhibition, additional information services. The admission charge is defined by three ticket levels: 8–10 EUR, 10–12 EUR, $>$ 12 EUR. When considering the location of the exhibition, we define a dichotomous attribute stating whether the exhibition is hosted in Venice or not. A further attribute distinguishes between information about the exhibition provided to visitors by the organizers and information gathered by visitors autonomously. Another dichotomous attribute refers to the presence (or absence) of additional multimedia services that contribute to the understanding of the exhibition. Starting from a full factorial which comprises $(2 \times 2 \times 2 \times 3 = 24)$ profiles, a fractional orthogonal factorial of eight profiles was created to capture main-effects of factors (attributes) Addelman (1962). Given this profile set, the existing status quo of services was not included in the choices set as a possible alternative.

4.2 Results

The summary part-worth utilities for each attribute level and the relative importance values assigned to the corresponding attributes are shown in Table 1. It emerges that visitors prefer Venice as a location for the exhibition rather than a different place. Visitors seem to be more interested in collecting information about the exhibit autonomously. In so doing, visitors show a preference for the provision of additional multimedia services which make the exhibition easier to understand. Table 1, Column 3 presents the relative importance for each attribute. We note that: the admission fee is the most important attribute in terms of relative importance; gathering information is the least important attribute; location and additional multimedia services show a similar level of relative importance. Given a change in the status quo, the estimated values shown in Table 1 can be used to derive an economic

Table 1 Part-worth utilities and attribute relative importance values. Venice. 2007

Attribute level	Part-worth utility	Attribute relative importance
Location: Venice	0.692	0.23526
Location: Other place	−0.692	//
Information concerning exhibition: Induced	−0.355	0.13948
Information concerning exhibition: Autonomous	0.355	//
Additional multimedia services: Present	0.567	0.24974
Additional multimedia services: Absent	−0.567	//
Admission charge: Ticket 8-10 EUR	1.266	0.37552
Admission charge: Ticket 11-12 EUR	−0.124	//
Admission charge: Ticket > 12 EUR	−1.143	//
Intercept	4.183	·

Table 2 Economic re-evaluation with the MI_j coefficient. Venice. 2007

Status quo	Modification of j attribute	MI_j	V_j
Venice (yes)	not	−0.06236	−5,811.84
Information concerning exhibition (induced)	autonomous	0.01899	1,770.01
Additional multimedia services (absent)	present	0.05430	5,061.19

estimation of revenue variation generated by that change in accordance with (6). First, we calculate the total utility associated with the status quo by summing the additive part-worth utilities of the corresponding attribute levels. Thus, we can suppose a change in the combination of levels and compute the total utility assigned to that alternative.

The first column of Table 2 reports the combination of attribute levels specifying the current service provision. If we suppose that the revenue generated by the existing status quo is 93,200 EUR (π), we can estimate the variation of revenue caused by changing one attribute level as shown in Table 2. Table 2 shows evidence that revenue would decrease if the exhibition were located in a different place ($-5,811.84$), whereas, revenue would increase by 5,061.19 EUR if the information were well-presented by using multimedia services. Moreover, visitors are more willing to pay for gathering information by themselves, therefore, the revenue gain is 1,770.01 EUR.

5 Conclusions

This work proposes a coefficient of economic re-evaluation based on conjoint analysis. A typical purpose of such analysis is to express a monetary evaluation related to a hypothetical change occurring in the combination of the attributes specifying the good. The proposed coefficient of economic re-evaluation works on part-worth utilities to determine which monetary valuation variations derive from the introduction of a change in the current specification of the good. Although conjoint analysis has been developed as a tool for market researchers to investigate the consumer preference structure, the suggested approach can be used for the valuation of non-market multi-attribute goods or services, such as cultural events. Results from a conjoint survey concerning a cultural event located in Venice reveal the way preferences affect the revenue generated by that cultural event. In conclusion, the article suggests an alternative approach to the willingness to pay measurement that exploits all the information collected in ranking or rating conjoint analysis response format. Moreover, our proposal has useful implications for the organisers of cultural events who can obtain information on revenue variation determinants. The further objective of this study is twofold. First, we aim at extending the suggested approach when a less general preference model than the part-worth one is used. Second, we argue that this technique may play a major role in investigating demand determinants if the respondent socio-economic characteristics are included in the analysis.

Acknowledgements We would like to thank "Fondazione di Venezia" for providing us with data used in our analysis.

References

Addelman, S. (1962). Symmetrical and asimmetrical fractional factorial plans. *Technometrics*, *4*(1), 47–58.

Boxall, P., Adamowicz, W., Swait, J., Williams, M., & Louviere, J. A. (1996). comparison of stated preference methods for environmental valuation. *Ecological Economics*, *18*, 243–253.

Carmone, F. J., Green, P. E., & Jain, A. K. (1978). Robustness of conjoint analysis: Some Monte Carlo results. *Journal of Marketing Research*, *15*(2), 300–303.

Cattin, P., & Wittink, D. R. (1982). Commercial use of conjoint analysis: A survey. *Journal of Marketing*, *46*(3), 44–53.

Chapman, R. G., & Staelin, R. (1982). Exploiting rank ordered choice set data within the stochastic utility model. *Journal of Marketing Research*, *19*(3), 288–301.

Desvousges, W., Smith, V. K., & McGivney, M. (1983). *A comparison of alternative approach for estimation of recreation and related benefits of water quality improvements* (Rep. No. EPA-230-05-83-001). Washington, DC: US Environmental Protection Agency.

Green, P. E., & Srinivasan, V. (1978). Conjoint analysis and consumer research: Issues and outlook. *Journal of Consumer Research*, *5*(2), 103–123.

Green, P. E., & Srinivasan, V. (1990). Conjoint analysis in marketing: New developments with implications for research and practice. *Journal of Marketing*, *54*(4), 3–19.

Irwin, J. R., Slovic, P., Licktenstein, S., & McClelland, G. (1993). Preference reversals and the measurement of enviromental values. *Journal of Risk Uncertain*, *6*, 5–18.

Mackenzie, J. (1993). A comparison of contingent preference models. *American Journal of Agricultural Economics*, *75*, 593–603.

Mazzanti, M. (2003). Discrete choice model and valuation experiment. *Journal of Economic Studies*, *30*(6), 584–604.

Placket, R. L. (1946). The design of optimum multifactorial experiments. *Biometrika*, *33*(4), 305–325.

Roe, B., Boyle, K. J., Teisl, M. F. (1996). Using conjoint analysis to derive estimates of compensating variation. *Journal of Environmental Economics and Management*, *31*, 145–159.

Sanz, J. A., Herrero, L. C., & Bedate, A. M. (2003). Contigent valuation and semiparametric methods: A case study of the National Museum of Sculpture in Valladolid, Spain. *Journal of Cultural Economics*, *27*, 241–257.

Srinivasan, V., & Shocker, A. D. (1973). Linear programming techniques for multidimensional analysis of preferences. *Psychometrika*, *38*, 337–369.

Stevens, T. H., Belkner R., Dennis D., Kittredge, D., & Willis, C. (2000). Comparison of contingent valuation and conjoint analysis in ecosystem management. *Ecological Economics*, *32*, 63–74.

Willis, K. G., & Snowball, J. D. (2009). Investigating how the attributes of live theatre production influence consumpion choices using conjoint analysis: The example of the National Art Festival, South Africa. *Journal of Cultural Economics*, *33*, 167–183.

Do Governments Effectively Stabilize Fuel Prices by Reducing Specific Taxes? Evidence from Italy

Marina Di Giacomo, Massimiliano Piacenza, and Gilberto Turati

Abstract After the sharp increase of oil prices experienced in recent years, in order to stabilize fuel prices, many countries experimented automatic fiscal mechanisms consisting of a one-to-one reduction in specific taxes matching the rise in input prices. This study investigates the impact of these mechanisms on wholesale gasoline and motor diesel prices. Our estimates highlight that fiscal sterilization brings about a rise in final wholesale prices that more than compensate reduction in taxes. Hence, these "flexible" taxation mechanisms could not be a proper policy for stabilizing price levels in fuel markets.

1 Motivation

As a reaction to oil price booms recorded in recent years, consumers' associations suggested (and policy makers experimented) the introduction of "flexible" taxation mechanisms on fuels. The idea of "flexible" taxation is very simple: in order to keep (gross) prices at a long-run equilibrium level, specific taxes should react one-to-one to variations observed in input prices. Indeed, among the various available measures, the sterilization of the increase in oil prices by a reduction in specific taxes on fuels seems to be one of the most popular actions.[1] However, such a sterilization policy should be carefully evaluated, as for the likely impact on consumers, producers, and tax revenues. On one side, if fuel prices are kept constant, there is a welfare enhancement for drivers and fuel consumers with respect to a situation of volatile prices. On

[1] Experiences of these policies can be found in different countries. In France, the government introduced the TIPP (the French specific tax on petroleum products) "flottante" in 2000, i.e., a fiscal mechanism able to change the tax in accordance with crude oil price trends. In Italy, two policy interventions were proposed in 2007 and 2008, but never implemented. Both of them envisaged some form of flexibility in the taxation mechanisms for fuels as a response to oil price peaks. Somewhat differently, in the U.S., a temporary tax moratorium on the sales tax was introduced by the Indiana and Illinois governors as a reaction to the gasoline price peaks during summer 2000 (Doyle and Samphantharak 2008).

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_11,

the other side, there is the need for the government to find different sources of tax revenues, or to reduce public expenditures. These concerns are particularly stringent in the European fuel markets, as fuel taxes account both for a large share of the retail price in many countries (particularly in Italy, where taxes represent more than 50% of the final consumer retail price), and for a nontrivial share of government's budget revenues (about 4–5% of total revenues), and finance both Central government and Local (Regional) governments expenditures.

The purpose of this brief note is to contribute to the current debate on fuel prices stabilization, by providing some insights on the possible effects of government strategies aimed at mitigating the impact of oil price peaks. We concentrate on the role of fuel specific taxes and estimate several reduced-form specifications considering as a dependent variable the equilibrium wholesale prices observed for both gasoline and motor diesel markets in Italy.

2 Data and Estimation Strategy

The main data source is the Bollettino Petrolifero (Oil Bulletin) published by the Italian Department for Economic Development. We collect data for three products: gasoline (unleaded and octave rating equal to 95 RON gasoline), motor diesel, and crude oil. For gasoline and motor diesel, we gather monthly data on wholesale prices, and the specific taxes over the period January 1996–December 2007, leading to time series of 144 observations each. We also obtain monthly C.I.F. (cost, insurance, and freight) crude oil prices for the same time period.

Similarly to other countries, the Italian fuel industry is characterized by a vertical structure involving three groups of actors: refiners, wholesale distributors, and downstream retailers. Refiners transform crude oil into petroleum products. Distributors receive petroleum products at their wholesale terminals and manage the distribution service to the gas stations. Finally, retailers sell products to final consumers. We concentrate on the segment where fuels (in our case unleaded gasoline and motor diesel) are delivered from the wholesale terminals to the retailers. The net wholesale price we observe is the equilibrium price in the market where distributors and retailers meet and includes distributors' profit margins, but it is net of specific and ad valorem taxes.

Table 1 shows some descriptive statistics for our sample. Wholesale prices for gasoline and diesel average 396 and 393 Euro per 1,000 liters, respectively. Diesel prices show some higher volatility than gasoline prices, but they are strongly correlated (correlation coefficient 0.96). Specific taxes amount on average to 615 Euro per 1,000 liters for gasoline and 452 Euro per 1,000 liters for motor diesel; they are lower for motor diesel over the whole sample period. On average, crude oil price over the 12 years is about 253 Euro per 1,000 liters, and the standard deviation is 164.

Our aim is to study the relationship between wholesale gasoline and motor diesel prices, on one side, and oil prices and specific taxes, on the other side. To this end, following the literature, we estimate a reduced-form specification, i.e., pricing equations where equilibrium prices are functions of exogenous demand, cost and market

Table 1 Sample descriptive statistics. Monthly observations on the Italian wholesale fuel market from 1996 to 2007

Variable description	Var. name	Mean	Std. dev.	25th pct.	Median	75th pct.
Gasoline price (Euro/1,000 liters)	PGAS	395.89	85.76	326.02	381.66	464.72
Diesel price (Euro/1,000 liters)	PDIES	392.67	111.38	295.30	361.98	487.31
Crude oil price (Euro/1,000 liters)	POIL	252.73	164.24	129.28	196.43	356.28
Gasoline specific tax (Euro/1,000 liters)	TAXGAS	614.86	31.80	592.00	601.87	649.72
Diesel specific tax (Euro/1,000 liters)	TAXDIES	452.13	22.93	431.86	445.16	475.07
Market share of leader firm (%)	LEADER	38.26	6.11	33.00	38.60	43.46
Share of population over 65 (%)	POP65	18.63	0.94	17.82	18.71	19.47
Number of vehicles/population ($\times$ 1,000)	VEHICLES	726.72	48.47	683.89	742.54	765.27
Number of retailers (10^3)	RETAIL	20.31	0.30	20.03	20.24	20.57
Quarterly per capita GDP (Euro)	GDP	6,156.8	302.2	5,914.8	6,268.0	6,373.0

Notes: All prices are deflated using the monthly Italian consumer price index (source Istat, base month: December 2007). The number of observations is 144.

power shifters (Chouinard and Perloff 2004, 2007, Alm et al. 2009). In particular we take to the data the following multiple time-series model:

$$\begin{aligned} PGAS_t &= \beta_0 + \beta_1 TAXGAS_t + \beta_2 POIL_t + X_t'\gamma + \epsilon_t \\ PDIES_t &= \alpha_0 + \alpha_1 TAXDIES_t + \alpha_2 POIL_t + X_t'\delta + \nu_t \end{aligned} \tag{1}$$

where the wholesale prices of gasoline (*PGAS*) and diesel (*PDIES*) are simultaneously regressed on a set of independent variables. TAX is the specific tax, different for gasoline (*TAXGAS*) and motor diesel (*TAXDIES*), *POIL* is the C.I.F. crude oil price,[2] while X is a vector collecting a set of additional covariates that we introduce to control for demand side and supply side factors that are common to both products. While coefficient on *POIL* is expected to be positive, we do not have clear a priori on the sign of the impact of specific taxes. As Hamilton (1999) shows, the sign of the coefficient is related to the elasticity of the demand schedule (see Di Giacomo

[2] We experimented with a number of empirical specifications that could account for the likely asymmetric response of fuel prices to oil price changes, often identified as "rockets" when oil prices go up, and as "feathers" when oil prices go down (e.g., Galeotti et al. 2003). We found only weak evidence of such asymmetry in price reactions. A possible explanation for this result is the monthly (instead of daily or weekly) frequency of the data.

et al. 2009, for more details on this point). In all specifications, we also include a set of monthly dummy variables, to capture seasonal effects in wholesale prices. With respect to the error terms, we assume that they are uncorrelated to the set of included regressors, while the contemporaneous errors can be correlated. We estimate the system of two equations in (1) by Zellner's (1962) seemingly unrelated regressions (SUR) estimator. The main advantage from this empirical strategy is a gain in efficiency with respect to the estimation of separate equations (Creel and Farell 1996). Before the estimation, all variables are transformed in natural logarithm, so as to allow for nonlinear relationships between fuel prices and the regressors.

The estimated relationships are *static* and should be interpreted as conveying information on *long run* behaviour of the variables of interests. Cointegration between the main variables (fuel prices, specific taxes and oil price) cannot be excluded,[3] and this suggests the existence of some sort of "equilibrium relationship" that is captured by the estimated reduced form model.

3 Results

Using our estimation results, we aim to give some insights first on the marginal effects of specific tax and oil price on gasoline and motor diesel wholesale prices, and then on the likely effects of sterilization policies. Table 2 shows three sets of different specifications for the model in (1). In the first set of estimates (MODEL 1), we report our basic specification, that includes only specific tax and oil price as explanatory variables.

In the second set of estimates (MODEL 2), we suspect the presence of some structural breaks that we ascertain by Chow break point test and CUSUM tests (Brown 1993). In particular, it is possible to single out two break points (one around the beginning of 2001, and the other at the beginning of 2004), which identifies three different sub-periods.[4] When specific tax and oil price are interacted with a set of three dummy variables, one for each of the three sub-periods characterizing our sample, we obtain that tax and oil elasticities are larger than those from the basic specification. Moreover, they sharply decreased during the second period (2001–2003), to return to original values in the last interval. The trend in the coefficients is likely to be associated to the scrutiny of the industry by the Italian Antitrust Authority, which was particularly severe at the beginning of 2000s (e.g., AGCM 2000). The reduction in elasticities in the second time period, especially with respect to oil price, may signal a change in the conduct by distributors that were under investigation (and successively fined) by the Antitrust Authority for the potential presence of a price cartel.

[3] Cointegration was tested using the Engle-Granger test (e.g., Koop 2007).

[4] The first sub-period, T1, is from January 1996 to December 2000, the second sub-period, T2, from January 2001 to December 2003, while sub-period T3 goes from January 2004 to December 2007.

Table 2 SUR estimation results: dependent variables are gasoline (PGAS) and diesel (PDIES) prices

	MODEL 1		MODEL 2		MODEL 3	
	PGAS	*PDIES*	*PGAS*	*PDIES*	*PGAS*	*PDIES*
TAX	−0.537***	−0.600***				
	(0.19)	(0.21)				
TAX_T1			−1.246***	−1.079***	−2.032***	−1.965***
			(0.24)	(0.27)	(0.40)	(0.29)
TAX_T2			−1.046***	−0.793***	−1.860***	−1.699***
			(0.24)	(0.27)	(0.40)	(0.29)
TAX_T3			−1.233***	−1.114***	−2.077***	−1.996***
			(0.24)	(0.28)	(0.40)	(0.30)
POIL	0.287***	0.395***				
	(0.02)	(0.02)				
POIL_T1			0.404***	0.472***	0.431***	0.488***
			(0.02)	(0.03)	(0.02)	(0.03)
POIL_T2			0.138***	0.124**	0.221***	0.183***
			(0.05)	(0.06)	(0.04)	(0.04)
POIL_T3			0.349***	0.480***	0.471***	0.511***
			(0.03)	(0.03)	(0.04)	(0.04)
LEADER					−0.384***	−1.138***
					(0.15)	(0.15)
POP65					−7.872***	−10.347***
					(1.81)	(1.82)
VEHICLES					29.648**	42.912***
					(11.96)	(12.16)
POP65_VEHICLE					−10.295**	−14.754***
					(4.19)	(4.26)
RETAIL					1.820**	1.658**
					(0.80)	(0.84)
GDP					1.134**	0.139
					(0.49)	(0.49)
Monthly dummies	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.99	0.99	0.99	0.99	0.99	0.99
Breusch-Pagan test of independence degrees of freedom [p-value]	94.93 1[0.00]		85.93 1 [0.00]		73.25 1 [0.00]	
Wald Statistic on TAX degrees of freedom [p-value]			25.25 3 [0.00]	39.60 3 [0.00]	28.60 3 [0.00]	50.94 3 [0.00]
Wald Statistic on POIL degrees of freedom [p-value]			29.35 3 [0.00]	39.89 3 [0.00]	28.57 3 [0.00]	50.19 3 [0.00]
N. of observations	144	144	144	144	144	144

Notes: The estimation method is Zellner's (1962) Seemingly Unrelated Regression (SUR). All variables have been transformed in natural logarithm. Standard errors are reported in round brackets. In MODEL 2 and MODEL 3 TAX and POIL have been interacted with three time dummies: T1, equal to one for observations from January 1996 to December 2000; T2 for observations from January 2001 to December 2003, and T3 from January 2004 to December 2007. The Wald Statistic on TAX tests the equality of the coefficients TAX_T1, TAX_T2 and TAX_T3. The Wald Statistic on POIL tests the equality of the coefficients POIL_T1, POIL_T2 and POIL_T3. * Significant at 10%. ** Significant at 5%. *** Significant at 1%.

Finally, MODEL 3 in Table 2 is an augmented specification that considers the role of exogenous supply side and demand side factors that may shift equilibrium prices (see Table 1 for a short description and descriptive statistics). We include the market share of the leader distributor in the Italian fuel industry (LEADER),[5] the share of population older than 65 (POP65), the number of vehicles per 1,000 inhabitants (VEHICLES) and its interaction with elderly people (POP65_VEHICLES), the number of gas stations (RETAIL), and per capita Gross Domestic Product (GDP). Overall, the new included variables are significant and exhibit the expected sign. All else equal, as the number of vehicles per capita increases, the prices rise. However, such a positive impact comes about at decreasing rates, for the effect of ageing population. A 1% increase in the number of gas stations is found to increase gasoline and motor diesel prices by approximately 1.7–1.8%, while the sign of GDP suggests a positive relationship between motor fuel prices and income, especially in the gasoline equation.[6] The coefficients for specific taxes are larger than in previous estimates, probably because of a better specification of the model.

We conduct a Wald test on the equality of tax and oil parameters across the three sub-periods for both MODEL 2 and MODEL 3. The hypothesis of equality is rejected by the data in both cases, supporting the existence of some structural breaks over the observed period (as shown by Chow and CUSUM tests). Depending on the sub-period being considered and the adopted specification, our results show that a 1% increase in oil price implies an increase of wholesale gasoline (diesel) prices ranging from 0.138 (0.124%) to 0.471 (0.511%). We also evaluate the incidence of specific taxes. Again, depending on the sub-period being considered and the chosen specification, we estimate that a 1% increase in the specific tax on gasoline is found to reduce wholesale gasoline price by 1.046–2.077%. For motor diesel, the effect of a 1% increase in the specific tax corresponds to a reduction in wholesale prices ranging between 0.793 and 1.996%.

We finally simulate the impact on wholesale prices of a sterilization policy that makes specific taxes react one-to-one to oil price increase relying on the richest specification (MODEL 3). Results, of course, differ according to the different sub-periods. Table 3 reports the predicted gasoline and diesel prices, with and without the policy intervention. In particular, we assess the effect of a 10 Euro increase in oil price sterilized by a 10 Euro reduction of specific taxes. Our evidence points to a positive impact of such a fiscal policy on fuel wholesale prices. In other words, no government policy would guarantee wholesale prices for gasoline (motor diesel) lower by around 3.10–3.46% (3.87–4.56%), depending on the sub-period being considered. These results are robust to a set of different checks (including the potential endogeneity of specific taxes and different forms of sterilization policies), as shown by Di Giacomo et al. (2009). Hence "sterilization" policies imply (at least partly) a direct transfer from the government to fuel distributors.

[5] The industry leader is ENI, whose main shareholder is the Italian Government.

[6] For a more detailed discussion of the impact exerted by supply and demand factors, see Di Giacomo et al. (2009).

Table 3 Fiscal policy simulation (evaluated at the sample mean values). Impact on wholesale gasoline and diesel predicted prices deriving from a 10 Euro decrease in the specific tax as a reaction to a 10 Euro increase in oil price

	PGAS			*PDIES*		
	T1	*T2*	*T3*	*T1*	*T2*	*T3*
Predicted price: no sterilization	434.41	407.85	409.05	448.77	418.20	424.11
(Euro/1,000 liters)	(9.49)	(11.42)	(10.90)	(10.04)	(11.90)	(11.61)
Predicted price: sterilization	449.13	420.48	423.22	468.94	434.40	443.47
(Euro/1,000 liters)	(10.35)	(11.15)	(12.00)	(11.21)	(11.78)	(11.91)
Absolute change (Euro/1,000 liters)	14.72	12.63	14.17	20.17	16.20	19.36
Percent change	3.39	3.10	3.46	4.49	3.87	4.56

Notes: Asymptotic standard errors are reported in round brackets. *PGAS* is the wholesale gasoline price, while *PDIES* is the wholesale diesel price. The "Predicted price: no sterilization" is the predicted wholesale fuel price, computed setting all variables at their sample mean values and increasing the mean oil price by 10 Euro: $\hat{P}_{no\text{-}steril.} = P(\overline{TAX}, \overline{POIL} + 10, \overline{X})$. The "Predicted price: sterilization" is the predicted wholesale fuel price, obtained by setting all variables at their mean values, increasing the mean oil price by 10 Euro, and subtracting 10 Euro from the mean specific tax value: $\hat{P}_{steril.} = P(\overline{TAX}-10, \overline{POIL}+10, \overline{X})$. Absolute change is the absolute difference between the predicted price with sterilization and the predicted price without sterilization. The Percent change is the relative (%) difference between the predicted price with sterilization and the predicted price without sterilization. Computations are based on results from MODEL 3 in Table 2. *T1* refers to sub-period T1 from January 1996 to December 2000, *T2* refers to sub-period T2 from January 2001 to December 2003, while *T3* is for sub-period T3, that goes from January 2004 to December 2007.

4 Concluding Remarks

As originated from the political debate following the peaks in oil price observed in recent years, the sterilization of oil price increase through a reduction in specific taxes seems to be one of the most popular measures. A complete evaluation of the welfare impact (on producers, consumers, State and regional finances) is clearly beyond the scope of our study, and our aim here is simply to give some insights on the likely effects of sterilization policies using our estimation results. Our estimates highlight that fiscal sterilization brings about a rise in final wholesale prices and suggest that "flexible" taxation mechanisms – focusing on specific tax reductions to compensate oil price increases – could not be a proper policy for stabilizing price levels in fuel markets.

References

AGCM, Autorità Garante della Concorrenza e del Mercato. (2000). *Accordi per la fornitura di carburanti*. Provvedimento n. 8353. Bollettino n. 22/2000. Available online at www.agcm.it.

Alm, J., Sennoga, E., & Skidmore, M. (2009). Perfect competition, urbanization, and tax incidence in the retail gasoline market. *Economic Inquiry*, *47*(1), 118–134.

Brown, R. L., Durbin, J., & Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society*, *B37*, 149–163.

Chouinard, H., & Perloff, J. M. (2004). Incidence of federal and state gasoline taxes. *Economics Letters*, *83*, 55–60.

Chouinard, H., & Perloff, J. M. (2007). Gasoline price differences: Taxes, pollution regulations, mergers, market power, and market conditions. *The B.E. Journal of Economic Analysis & Policy*, *7*(1), Article 8, available at: http://www.bepress.com/bejeap/vol7/iss1/art8

Creel, M., & Farell, M. (1996). SUR estimation of multiple time-series models with heteroscedasticity and serial correlation of unknown form. *Economics letters*, *53*, 239–45.

Di Giacomo, M., Piacenza, M., & Turati, G. (2009). *Are "Flexible" taxation mechanisms effective in stabilizing fuel prices? An evaluation considering the Italian fuel markets*. University of Torino, Department of Economics and Public Finance "G. Prato", Working Paper n. 7, available at: http://ideas.repec.org/p/tur/wpaper/7.html

Doyle, J., & Samphantharak, K. (2008). $2.00 Gas! Studying the effects of a gas tax moratorium. *Journal of Public Economics*, *92*, 869–884.

Galeotti, M., Lanza, A., & Manera, M. (2003). Rockets and feathers revisited: An international comparison on European gasoline markets. *Energy Economics*, *25*, 175–190.

Hamilton, S. (1999). Tax incidence under oligopoly: A comparison of policy approaches. *Journal of Public Economics*, *71*, 233–245.

Koop, G. (2007). *An introduction to econometrics*. New York: John Wiley and Sons.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, *57*, 348–68.

An Analysis of Industry Sector Via Model Based Clustering

Carmen Cutugno

Abstract The paper presents an unsupervised procedure for the evaluation of the firm financial status, aiming at identifying a potentially weak level of solvency of a company through its positioning in a segmented sector. Model Based Clustering is, here, used to segment real datasets concerning sectoral samples of industrial companies listed in five European stock exchange markets.

1 Introduction

The evaluation of a firm financial status depends on many factors, financial and economic, that could be predictive of a weak level of solvency or a pre-condition for a liquidation proceeding, stating the inability of the firm to pay its financial obligations. The information about the financial status of a firm and its positioning in the belonging industry-sector is crucial for investors, stockholders, loan-analysts or creditors, taking into concern the description and the representative characteristics of the analysed sector.

Several estimation methods have been suggested, in literature, to predict financial distress, from the simple univariate analysis (Beaver 1966), to multiple discriminant analysis (MDA) (Altman 1968, Taffler 1982), logit (Ohlson 1980) and probit models (Zmijewski 1984), artificial neural network models (ANN) (Tam and Kiang 1992, Chen and Du 2009), rough set theory (Dimitras et al. 1999), Bayesian network (BN) models (Sun and Shenoy 2007), and genetic programming (Lensberg et al. 2006). The above-mentioned methodologies are considered as supervised classification methods, where a collection of labelled patterns are provided and the problem is to label a new unlabelled item. The historical training patterns are used to learn and to derive rules of classes (Jain et al. 1999). Supervised learning can, usually, achieve high prediction accuracy if the training data and the analysed data have similar characteristics. The disadvantage emerges if there are no data records with known class labels. In bankruptcy studies, samples tend to be small or the information about the failure status may not be readily available for training. Differently, data mining techniques and clustering methods belong to the unsupervised classification methods

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_12,

dealing with the problem of predicting unobserved features (cluster labels) from observed ones, so category labels are data driven.

Recent researches have proposed the application of clustering analysis and data mining in the field of the performance evaluation of industrial organization (Azadeh et al. 2007, Rezaie et al. 2009) and financial distress prediction (Bensmail and De Gennaro 2004, Sun and Li 2008).

In the literature, the variables generally considered in the analysis concerning distress predicting models consist of financial ratios or a set of financial and economic ratios, or sometimes different variable classes composed of both ratios and balance-sheet items.

The underlying idea in this paper is that the financial and economic features, defining the financial structure level, are strictly connected with the industry sector the firm belongs to Gupta and Huefner (1972), and that seeking for industry-sector key indicators levels may lead to a more appropriate evaluation of the firm financial profile.

In the paper, an unsupervised procedure of classification analysis is presented, aiming at identifying the position of a company in a segmented sector. The choice of a data-driven methodology of analysis is due to the consideration that the information about a liquidation proceeding does not identify potential failure pre-condition because it refers to a stated insolvency status.

The time period considered in the evaluation process is from 2005 to 2007, in order to verify the classification dynamic of a firm in the sectoral segmented framework.

The procedure presented starts from the assessment of the financial and economic indicators that influence the specific industry sector, by a principal component analysis (PCA), and then it proceeds with operating the segmentation, in a financial and economic perspective, of the sector by clustering methodology.

We propose the use of the model based clustering because it allows the modelization of the covariance matrix and because of its capability to assign every unit to a n-group with a probability of belonging. The model based methodology is compared to another clustering method, the K-means clustering. In model based clustering, it is assumed that the data are generated by a mixture of underlying distributions in which each component represents a different group or cluster and the problem of determining the number of clusters and of choosing an appropriate clustering method can be recast as statistical model choice problems (Banfield and Raftery 1993, Fraley and Raftery 2002).

The rest of the paper is organized as follows, a brief review about the model based clustering methodology, then the presentation of two case-studies on two different industrial sectors, Constructions and Technology Hardware, and the results of the proposed procedure.

2 Basic Ideas on Model Based Clustering

In model based clustering, each cluster is, generally, represented by a Gaussian model:

$$\phi_j(x|\mu_j, \Sigma_j) = (2\pi)^{-\frac{p}{2}} |\Sigma_j|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x_i - \mu_j)^T \Sigma_j^{-1}(x_i - \mu_j)\right\}, \quad (1)$$

where x represents the data, and j is an integer subscript specifying a particular cluster. Formerly, the mixture models for clustering analysis considered only equal covariance matrix Σ. Model-based clustering offers different modelization of the covariance matrix Σ, that could be parametrized by spectral decomposition, in the form:

$$\Sigma_j = \lambda_j D_j A_j D_j', \quad (2)$$

where $\lambda_j = |\Sigma_j|^{\frac{1}{q}}$, D_j is the orthogonal matrix of eigenvectors of Σ_j and A_j is a diagonal matrix whose elements are proportional to the eigenvalues of Σ_j (Banfield and Raftery 1993). The orientation of the principal components of Σ_j is determined by D_j, while A_j determines the shape of the density contours; λ_j specifies the volume of the corresponding ellipsoid, which is proportional to $\lambda_j^d |A_j|$, where d is the data dimension. Characteristics (orientation, volume and shape) can vary between clusters, or be constrained to be the same across clusters. In this case, clusters are ellipsoidal, centred at the means μ_j. The covariances Σ_j determine their other geometric features. In the model based clustering, we assume a mixture model with incomplete data and we use the EM algorithm (Dempster et al. 1977). EM algorithm maximizes the likelihood function $L(\psi|x_1, \ldots, x_n)$ indirectly by proceeding iteratively in two steps, E-step and M-step, applied on the complete data log likelihood function, $\log L_c(\psi)$. EM is strongly sensitive to initialization, being a local maximizer seeker, and also, because of the unboundiness of the likelihood function, the optimization could fail, converging to some singularities, Ingrassia and Rocci (2007) for constrained ML formulations. A procedure is to initialize EM with the model based hierarchical results and to use approximate Bayes factors with the BIC (Bayes Information Criterion) to determine the number of clusters, see Fraley and Raftery (2002).

3 Two Case-Studies

We consider two datasets of yearly financial statement data of companies, selected according to the Industry Classification Benchmark (ICB), listed on Frankfurt, Madrid, Paris, London and Milan stock exchange markets, for the period 2005–2007. The first sample refers to the Technology Hardware sector and consists of 92 units for the period 2005–2007. The second sample refers to the Constructions sector and consists of 105 units for 2005, 107 units for 2006 and 113 units for 2007. We calculate a set of 13 financial and economic ratios: Quick ratio, Current ratio, Leverage, Debt Coverage, Cost of Debt, Equity to liabilities, Asset turnover, Expected time of liabilities refunding indicator, Ebitda to Sales ratio, Return on Asset (Roa), Return on Investment (Roi), Return on Sales (Ros) and Return on Equity (Roe).

3.1 The Statistical Analysis and Results

Firstly, we process a principal component analysis (PCA), as a pre-step examination on the variables and their influence on the data variability, then we run a model based clustering on the scores obtained by the PCA, in order to classify the companies into groups related to different financial structure levels. The variables (financial and economic ratios) are calculated by operating transformations of accounting data measured in the same unit. We do not standardize the variables to compute the principal components because measurements are on comparable scale, see Jo et al. (1997), Lee et al. (2005). The clustering analysis has been processed by using the package MCLUST vers.3.3.1 (Fraley and Raftery 2009) of the statistical software R. We select the best model according to the BIC corresponding to the different parametrisation of the covariance matrix Σ_j, and indicating the number of the component of the mixture. Each unit is assigned to the component to which it has the highest estimated posterior probability of belonging and each distribution component of the mixture may correspond to a cluster and thus, in our analysis, to a group of companies (Fraley and Raftery 2002).

By examining the cluster centroids, mean and median, of the ratios, we may define different financial and economic structure levels in each component of the mixture. For the first dataset, Technology Hardware sector, we select, by the PCA, three components explaining about the 72% of variance in 2005, 66% in 2006 and 68% in 2007. In all the three years, 2005–2007, the components extracted are strongly influenced by the evaluation of working capital, as expressed by the operating return and the relation between short-term debts and current assets, and also by the evaluation of the weight of net equity. In 2005 and 2007, we found a strong influence on the evaluation of the economic return for investors (or the ownership), expressed by the Roe, less strong in 2006. The evaluation of the firm's ability to refund financial debts and on the expense on debts, expressed by Debt Coverage and Cost of Debt, is, strongly, captured by the components extracted in the whole period. The asset turnover is relevant in 2005 and 2006, less in 2007. The operating return evaluation, is strongly captured in the whole period. In 2005, by the application of the model based clustering on the scores, as shown in Fig. 1, the data have been fitted by a three-components mixture of Gaussian distributions, connected with a VEV,three model, thus a ellipsoidal, equal shape model, presenting cluster 1 with about 53% of the observations, cluster 2 with 36% and cluster 3 with 11%, also see Table 1. Cluster 1 presents high levels of turnover, economic returns and indebtedness, see Table 2. Cluster 2 shows a medium level of indebtedness, low operating returns and asset turnover. Cluster 3, a marginal group, presents high level of indebtedness and low operating returns. In 2006, the data have been fitted by a four-components mixture of Gaussian distributions, connected with a VEV, four model, thus an ellipsoidal, equal shape model. Cluster 1 with about 37% of the observations, cluster 2 with 23%, cluster 3 with 28%, and cluster 4 with 11%. Cluster 1 presents high economic return levels and quite high level of indebtedness. Cluster 2 presents a lower level of indebtedness, compared to group 1, but a very low level of operating economic return. Cluster 3 is highly indebted, not very

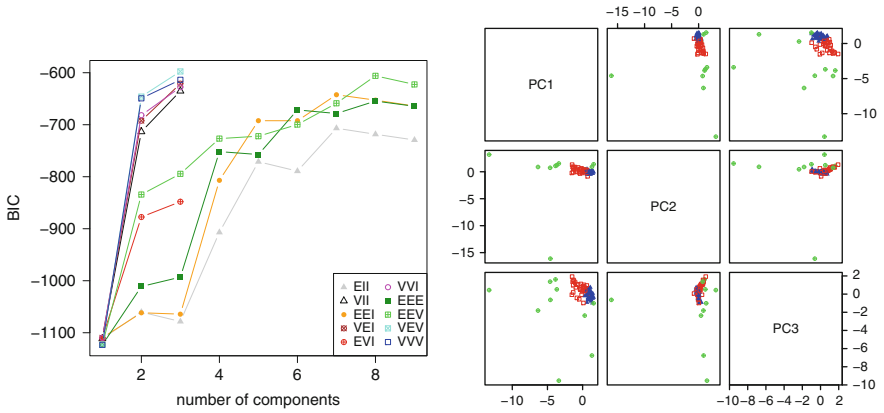


Fig. 1 Model based clustering for technology hardware sector, year 2005: BIC values and model selection; Pairs plot of model based classification of the data

Table 1 Model and number of clusters selected

Year	Constructions sector	Technology hardware sector
2005	VVI,3	VEV,3
2006	VEI,4	VEV,4
2007	VEV,3	VEV,4

Table 2 Industry sector analysis

Technology hardware	Cluster 1	Cluster 2	Cluster 3	Cluster 4
2005	High econ. returns	Low econ. returns-	Low econ. returns- -	
	High indebtedness	Medium indebt.	High indebtedness+	
	High turnover	Low turnover	Low turnover	
2006	High econ. returns	Low econ.returns-	Low econ. returns	Low econ. returns-
	High indebtedness	Medium indebt.	High indebtedness	High indebtedness
	High turnover	Medium turnover	Medium turnover	Low turnover
2007	Medium econ. ret.	Medium econ. ret.	Low econ. returns-	Low econ. returns-
	Medium indebt.	Low indebtedness	High indebtedness	Not coherent
	Medium turnover	Medium turnover	High turnover	
Constructions	Cluster 1	Cluster 2	Cluster 3	Cluster 4
2005	High econ. returns	Low econ. returns	Medium econ. ret.	
	High indebtedness–	High indebt.++	High indebtedness	
2006	High econ. returns	Medium econ. ret.	Low econ. returns	Low econ. returns–
	Medium indebt.	Medium indebt.	High indebtedness	High indebt.++
2007	Low econ.returns	High econ. returns	Medium econ. ret.	
	High indebt.++	Medium indebt.	Medium indebt.	
	Low refund. capab.-	Medium ref. capab.	Low refund. capab.	

high economic return levels, even if with a current financial management better than cluster 2. Cluster 4, a residual group. In 2007, the data have been fitted by a four-components mixture of Gaussian distributions, connected with a VEV,four model, thus an ellipsoidal, equal shape model. Cluster 1 with about 48% of the observations, cluster 2 with 27%, cluster 3 with 17%, and cluster 4 with 7%. Cluster 1 presents an average level of indebtedness and asset turnover, with economic return levels not very high, connected with a quite high level of cost of debt. Cluster 2 shows low economic return levels, low indebtedness, medium levels of asset turnover, and a quite good current financial situation. Cluster 3 presents good levels of asset turnover, but quite high levels of indebtedness and low economic return levels. Cluster 4 is a marginal group, presenting group centroids not very coherent from the economic point of view.

For the second dataset, Construction sector, we select, by the PCA, three components explaining about the 69% of variance in 2005, 76% in 2006 and 79% in 2007. In all the three years, 2005–2007, the components extracted are, strongly, influenced by the evaluation of working capital, as expressed by the operating return and the relation between short-term debts and current assets, and also by the evaluation of the weight of net equity. In 2006 and 2007, we found a strong influence on the evaluation of the economic return for investors, expressed by the Roe, less strong in 2005. The evaluation of the firm's ability to refund financial debts, expressed by Debt Coverage and Cost of Debt, is captured by the components extracted in the whole period. The asset turnover is relevant in 2005, less in 2006 and 2007.

In 2005, the data have been fitted by a three-components mixture of Gaussian distributions, connected with a VVI,3 model, thus a diagonal, varying volume and shape model. Cluster 1 with about 56% of the observations, cluster 2 with 13% and cluster 3 with 31%. Cluster 1 presents low Asset Turnover and an high indebtedness, even if better levels of economic return. Cluster 2 shows an high level of indebtedness with low level of turnover and economic return. Cluster 3 presents average levels of economic return and a good level of turnover, even if characterized by high indebtedness level. In 2006, the data have been fitted by a four-components mixture of Gaussian distributions, connected with a VEI, four model, thus a diagonal, equal shape model. Cluster 1 with about 23% of the observations, cluster 2 with 63%, cluster 3 with 7%, and cluster 4 with 7%. Cluster 1 presents good levels of economic return and an average level of indebtedness. Cluster 2 shows economic returns and indebtedness levels on average. Both cluster 3 and cluster 4 represent marginal groups with few elements. In 2007, the data have been fitted by a three-components mixture of Gaussian distributions, connected with a VEV, three model, thus an ellipsoidal, equal shape model. Cluster 1 with about 5% of the observations, cluster 2 with 41% and cluster 3 with 54%. Cluster 1, a marginal group, with high leverage. Cluster 2 shows medium level of indebtedness and high economic returns. Cluster 3 presents an average economic and financial situation.

The identification, in both three years, of marginal groups with dispersed elements, could be interpreted as a signal of the presence of potential outliers. These findings have been detected both in the first and the second dataset.

In order to compare the methodology presented in the paper, we have processed on the two datasets for the three years, 2005–2007, a K-means clustering, and we found, for all the runs, classifications dissimilar to the ones obtained in the model based clustering. We observed that the model based methodology detects more clusters and one or two residual clusters compared to the K-means procedure that often tends to identify fewer larger cluster and some singletons.

4 Conclusions and Further Developments

In this part of a more extended company analysis framework, we have considered a two-ways data, and we have followed a sequential procedure by applying firstly the PCA and then the clustering to the scores, a procedure usually used in literature. Our intent is both to extend the analysis to other industrial sectors and to consider a different procedure, consisting in the simultaneous combination of dimensionality reduction and clustering operation, see Vichi et al. (2007).

The results may suggest that model based clustering is a more flexible methodology, compared to the other clustering methods, because every unit is assigned to every of the n-group with a probability of belonging (the posterior probability), giving the possibility to better identify borderline unit. In our application, on average, we find that model based clustering tends to detect more clusters than K-means. Similar findings have been, also, found in previous papers, see Atkinson et al. (2010), where MCLUST is compared to another robust clustering method.

The identification of the number of the Gaussian components of the mixture with the number of clusters may need further analysis in order to verify possible misleading association, signalled by the presence of components with few elements, or units with not very high posterior probability of belonging and not very well separated groups, that could be connected with the merging problem of normal distributions or not Gaussian distributions. Moreover, both dispersed few elements group or very low probability of belonging of an element to a cohesive group could indicate the presence of potential outliers.

We intend to proceed with further research in order to provide a more robust model based approach for clustering, by considering mixture of t distributions instead of Gaussian mixture, see Peel and McLachlan (2000) or other robust clustering methodology.

References

Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609.

Atkinson, A. C., Riani, M., & Cerioli, A. (2010). Robust clustering for performance evaluation. In F. Palumbo et al. (eds.), *Data analysis and classification* (pp. 381–390). Berlin Heidelberg: Springer-Verlag.

Azadeh, A., Ghaderi, S. F., Miran, Y. P., Ebrahimipour, V., & Suzuki, K. (2007). An integrated framework for continuous assessment and improvement of manufacturing system. *Applied Mathematics and Computation*, *186*, 1216–1233.

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, *49*(3), 803–821.

Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research, Empirical Research in Accounting: Selected Studies*, *4*, 71–111.

Bensmail, H., & DeGennaro, R. P. (2004). Analyzing Imputed Financial Data: A New Approach to Cluster Analysis. *FRB of Atlanta Working Paper No. 2004-20*. Available at SSRN: http://ssrn.com/abstract=594383

Chen, W. S., & Du, Y. K. (2009). Using neural networks and data mining techniques for the financial distress prediction models. *Expert system with Application*, *36*, 4075–4086.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Royal Statistical Society. Series B, 39*(1), 1–38.

Dimitras, A. I., Slowinski, R., Susmaga, R., & Zopounidis, C. (1999). Business failure prediction using rough sets. *European Journal of Operational Research*, *114*, 263–280.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, *97*(458), 611–631.

Fraley, C. & Raftery, A. E. (2006, revised 2009). Mclust version 3 for R: Normal mixture modeling and model-based clustering. Technical Report no. 504, Department of Statistic, University of Washington. Statistic, University of Washington.

Gupta, M. C., & Huefner, R. J. (1972). A cluster analysis study of financial ratios and industry characteristics. *Journal of Accounting Research*, *10*(1), 77–95.

Ingrassia, S., & Rocci, R. (2007). Constrained Monotone EM Algorithms for finite mixtures of multivariate Gaussians. *Computational Statistics and Data Analysis*, *51*, 5339–5351.

Jain, A. K., Murthy, M. N., & Flinn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, *31*(3), 264–323.

Jo, H., Han, I., & Lee, H. (1997). Bankruptcy prediction using case-based reasoning, neural networks, and discriminant analysis. *Expert System with Application*, *13*(2), 97–108.

Lee, K., Booth, D., & Alam, P. (2005). A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms. *Expert system with Application*, *29*, 1–16.

Lensberg, T., Eilifsen, A., Mckeee, T. E. (2006). Bankruptcy theory development and classification via genetic programming: Some methodological issues. *European Journal of Operational Research*, *169*, 677–697.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, *18*(1), 109–131.

Peel, D., & McLachlan, G. J. (2000). Robust mixture modelling using the distribution. *Statistics and Computing 10*, 339–348.

Sun, J., & Li, H. (2008). Data mining method for listed companies financial distress prediction. *Knowledge-Based Systems*, *21*, 1–5.

Sun, L., & Shenoy, P. P. (2007). Using bayesian networks for bankruptcy prediction: Some methodological issues. *European Journal of Operational Research*, *180*(2), 738–753.

Taffler, R. J. (1982). Forecasting company failure in the UK using discriminant analysis and financial ratio data. *Journal of the Royal Statistical Society. Series A*, *145*(3), 342–358.

Tam, K. Y., & Kiang, M. Y. (1992). Application of neural networks: The case of bank failure predictions. *Management Science*, *38*(7), 926–947.

Rezaie, K., Dehghanbaghi, M., & Ebrahimipour, V. (2009). Performance evaluation of manufacturing systems based on dependability management indicators-case study:chemical industry. *International Journal of Manufacturing Technology & Management*, *43*, 608–619.

Vichi, M., Rocci, R., & Kiers, H. A. L. (2007). Simultaneous component and clustering models for three-way data: within and between approaches. *Journal of Classification*, *24*, 71–98.

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, *22*, 59–82.

Impact of Exogenous Shocks on Oil Product Market Prices

Antonio Angelo Romano and Giuseppe Scandurra

Abstract The presence in Italy of a high number of vertically integrated energy companies, has given us the idea to investigate the effects that adoption of new price policies, and geopolitical events, have on the mechanisms of price transmissions in the Italian wholesale and retail gasoline markets, using weekly data from January 03/2000 to November 28/2008. The interaction among crude oil prices, gasoline spot prices and the before tax gasoline retail prices have been considered. The results show that industrial policies have a significant role in explaining gasoline prices. To be more specific, shock in the retail market has an important role in the increasing price of gasoline.

1 Introduction

Since May 1994 the Italian gasoline prices have been established by the distributing companies, that determine the final price and recommend it to the managers of the single selling points. We would like to briefly summarize the gasoline price composition. The recommended final price is determined by the sum of the gasoline industrial price and the fiscal component (excise tax and VAT). The industrial price includes all the costs of the material, refinement, storage, primary and secondary distribution and other structural costs (logistic, commercial, administrative, managing agent margins, as well as industrial margins). The international quotation of the refined product represents about 33% of the gasoline price and it is also the primary component of the industrial price; for Italy it is the *Platt's Cif Med*.[1] Gasoline prices depend on the dynamics of the international quotations of crude oil and on the productive capacity of the refineries. Considering the close relationship that characterizes the gasoline markets it is useful to analyze them from a double point of view. We consider two independent markets, as suggested by Unione Petrolifera.

[1] Platt's Cif Med points out the cost of the importation in the Mediterranean zone of oil products and it is used by the oil companies as proxy of the refinement costs.

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_13,

The wholesale market considers only crude oil and Platt's prices while the retail market takes into account Platt's price and before-tax gasoline price. In the two stages of the production–distribution chain we think that a shock in the markets has consequences on the prices of oil by-products.

A great number of studies emphasize the effects that exogenous shocks have on oil market. In fact, existing literature contains several attempts to identify the effects of structural shocks on certain macroeconomic variables, such as, real GDP, inflation, employment (see e.g., Hamilton 1983, Hamilton and Herrera 2002, Hamilton 2003, Rodriguez 2008). Fewer research have been carried out to investigate the effects of oil price changes on the asset price (see e.g., Jones and Kaul 1996, Apergis and Miller 2009). However, we find that a less explored aspect is related to the effects that exogenous shocks have on the market mechanism. One important issue is connected to the impact that adoption of new industrial policies and/or geopolitical events have on the oil product prices. In wholesale market, *Energy Information Agency* (EIA) identifies about 15 relevant events in the oil market price for most of the period (January, 2000–December, 2007) taken in consideration. To these, we should add some speculative movements of the futures markets that are not being considered in this work. Furthermore, ENI, the Italian market leader in the gasoline distribution, has adopted a new price policy in the retail market. It has broken the common practice to follow Platt's quotation to adjust gasoline prices with frequent but small variations. From October 2004, ENI's new price policy has foreseen a frequent loss with more consistent price changes. Many companies operating in the Italian retail gasoline market have followed this new policy, except for Erg. These choices have dropped gasoline industrial price to short run variation of Platt's quotations (Autorita' Garante della Concorrenza e del Mercato 2007).

Aim of this paper is to assess the impact that exogenous shocks have on oil product prices. We estimate the relationship at different stages of production–distribution chain. In order to have the impact effect of the new price policy, in this work we separately estimate the relationship between (i) oil price and Platt's quotation (wholesale market) and between (ii) Platt's and gasoline prices (retail market). The choice is justified by the fact that a company like ENI affects both the relationship between oil price in wholesale market, because it is a integrated company, and the relationship between gasoline prices in the retail market, because it is also a distribution company. The model we estimate is a Vector Error Correction (VECM) with intervention variables.

The organization of the paper is as follow. Section 2 describes data. In Section 3 we analyze the impact of exogenous shocks in the two markets. Section 4 closes with some concluding remarks.

2 Data, Integration Order and Structural Break Tests

Data are the weekly time series from January 03/2000 to November 28/2008 of crude oil price (Brent), gasoline spot price (Platt's) and the before-tax gasoline retail price (Gasoline). In particular, we consider the Europe Spot Price, available from

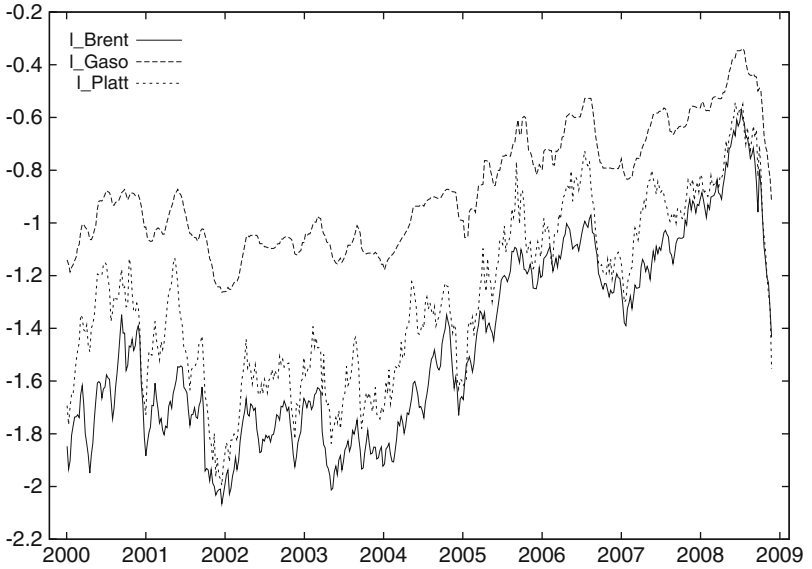


Fig. 1 Crude oil price, Platt's quotation and net-of-taxes gasoline price. January 03, 2000–November 28, 2008

EIA; for the ex-refinery gasoline price we consider *Platt's CIF/Med*. The net-of-taxes retail price is from *Italian Ministry of Economic Development*. All prices are expressed in logarithmic (Fig. 1). We use the before-tax gasoline price because of the different taxation gasoline has had in the period in analysis. Furthermore, in order to orthogonalize the effect of different exchange rate, we convert crude oil and Platt's original prices, that are expressed in \$/mton into €/lt prices (Apergis and Miller 2009). So, these prices are converted using the proper conversion coefficients. For the conversion of US dollar into euro price, we use the weekly exchange rates available on the web site of the Bank of Italy. In so doing we take into account the favorable exchange rate for Italian firms in the last years. In order to test the impact of exogenous shocks in price formation in the production–distribution chain, we have to test for the presence of a stochastic trend in the autoregressive representation of each individual series with Augmented Dickey Fuller (ADF) and Phillips-Perron (PP) tests. However, in a seminal paper, Perron (1989) showed that a break in the deterministic time trend dramatically reduces the power of standard unit root tests because the possibility of a break changes the (asymptotic) distribution of the test.

Similarly, Gregory et al. (1996) demonstrated that the rejection frequency of the ADF test for cointegration falls substantially in the presence of a structural break in the cointegrating relation. A break, in fact, introduces spurious unit root behavior in the cointegrating relationship so that the hypothesis of no cointegration is difficult to reject. Given the presence of shocks in markets, as ENI's policy and EIA's oil

Table 1 Minimum t statistics for Zivot-Andrews in the three time series and in presence of break in intercept, trend or both

	Intercept			Trend			Both		
	Lag	Min. t-stat	Critical value at 5%	Lag	Min. t-stat	Critical value at 5%	Lag	Min. t-stat	Critical value at 5%
Brent	4	−3.661	−4.80	1	−3.688	−4.42	1	−3.919	−5.08
Platt's	3	−4.044	−4.80	3	−3.851	−4.42	3	−4.091	−5.08
Gasoline	1	−3.289	−4.80	1	−2.933	−4.42	1	−3.131	−5.08

market shocks, we firstly propose the Zivot-Andrews[2] (ZA) unit root test (Zivot and Andrews 1992). The null hypothesis is that the series follow a random walk process without structural change, while the alternative is that the series are trend stationary with one-time break with the precise timing unknown (e.g., Zivot and Andrews 1992, Hondroyiannis 2004). The Zivot-Andrews [3] test is estimated for the variables used in the analysis to ensure that all series are *I(1)*. Table 1 reports the *minimum t-statistics* from testing the stationarity assuming a shift in mean, in trend or both for the first differences of crude oil price, Platt's quotation and before-tax gasoline retail price. The *minimum t-statistics* reported are the minimum overall break point regressions from January 03/2000 to November 28/2008.

The results suggest that at 5% level of significance none of the estimated variables are stationary around a broken trend or a shift in the mean, while their difference is *I(0)*.The Zivot-Andrews test confirms that all the variables are *I(1)*, since the previous test confirms the stationarity of the first differences of the variables at different levels of significance.

However we report also ADF[4] and Phillips-Perron (PP) tests, cf. Table 2. ADF and PP tests fail to reject null hypotheses for oil price, Platt's quotation and net-of-tax gasoline price. We have come to the conclusion that these variables have stochastic trends. We now test for cointegration in the two markets, applying the Gregory-Hansen (1996) procedure,[5] cf. Table 3.

The Gregory and Hansen advanced three models including model C, model C/T and model C/S, also known as 'regime shift' model. Model C allows for a level shift.

[2] The break date is selected where the t-statistic from the ADF test of unit root is at a minimum (most negative). The critical values in Zivot and Andrews (1992) are different to the critical values in Perron (1989). The difference is due to that the selecting of the time of the break is treated as the outcome of an estimation procedure, rather than predetermined exogenously.

[3] The Zivot-Andrews test has been recently criticized because it assumed that if a break occurs, it does so only under the alternative hypothesis of stationarity. This is undesirable since (i) it imposes an asymmetric treatment when allowing for a break, so that the test may reject when the noise is integrated but the trend is changing; (ii) if a break is present, this information is not exploited to improve the power of the test (Kim and Perron 2009).

[4] The optimal lag length is taken to be the one selected by Akaike Information Criterion (AIC).

[5] We also test cointegration hypothesis with the Engle-Granger (2009) two step procedure and we confirm it.

Table 2 Augmented Dickey Fuller and Phillips-Perron unit root test results for oil price, Platt's quotation and net-of-tax gasoline price

Augmented Dickey Fuller test	Lag	T stat	p-value	Lag	T stat	p-value	Decision
Brent	1	−1,978	0,297	0	−16,688	0,000	I(1)
Platt's	0	−2,126	0,235	0	−21,370	0,000	I(1)
Gasoline	4	−2,261	0,185	6	−7,180	0,000	I(1)
Phillip Perron test		T stat	p-value	Lag	T stat	p-value	Decision
Brent		−1,717	0,422		−16,594	0,000	I(1)
Platt's		−2,303	0,172		−21,442	0,000	I(1)
Gasoline		−1,896	0,334		−13,430	0,000	I(1)

Table 3 Gregory-Hansen cointegration test

Model	Wholesale market	Retail market
C	−5,47[a]	−6,61[a]
C/T	−6,04[a]	−7,50[a]
C/S	−5,46[b]	−6,94[a]

[a] significant at 1%.
[b] significant at 5%.

Model C/T allows for a level shift and a time trend. Model C/S allows for a shift in both intercept and slope. The cointegration tests of Gregory and Hansen (1996) explicitly extend the Engle-Granger test to allow for potential structural change in the cointegrating relationship. We reject the null hypothesis of no cointegration. Oil price and Platt's quotation (wholesale market) share a stochastic trends and a long-run equilibrium relationship exists among these price. The same is valid for retail market (Platt's quotation and net-of-tax gasoline price).

3 The Impact of Exogenous Shocks

Although the ZA test indicates empirical evidence about the presence of a unit root in the series (without shock), the presence of exogenous shocks singled out by EIA and the knowledge of a new price policy adopted by ENI has given us the idea to analyze if an economic impact of these on oil product prices in retail and wholesale markets really exists. In order to test the effects that geopolitical events and/or industrial policies have had on mechanism of price transmission in Italian wholesale and retail markets we propose the use of a Vector Error Correction Model (VECM) with intervention variables (Luktepol 2005). Furthermore, a VECM allow us to test a feedback between variables in the production–distribution chain. We estimate the

following model for each market stage[6]:

$$\Delta X_t = \delta + \Psi X_{t-1} + \sum_{i=1}^{p-1} \Phi \Delta X_{t-1} + D\mathscr{Z}_{k,t} + \epsilon_t \tag{1}$$

where:
δ is the vector of constants, X_t is the matrix of endogenous variables (i.e., oil and Platt's prices for the wholesale market; Platt's quotations and net-of-tax gasoline prices for the retail market), $\mathscr{Z}_{k,t}$ is the vector of the different intervention variables. Moreover, Ψ is a reduced rank $r < 2$ coefficients matrix that can be decomposed into two vectors, α and β, and such that $\Psi = \alpha\beta'$ is stationary. The exogenous shocks should now be identified in the two market stages. In the wholesale market different intervention functions are taken in consideration. As described in Sect. 1, EIA singles out about 15 relevant events for most of the period (January, 2000–December, 2007) examined. To these, probably some speculative movements in the futures market that are not considered in this work should be added. However, we believe that not all the events recognized by EIA have a straight impact on the series. As example, an estimate of the effect of crude oil price decline after September 11/2001 and with the military action in Iraq beginning on January 07/2002 are reported. The step functions used in the wholesale market are the following:

$$\mathscr{Z}_{Terrorist,t} = \begin{cases} 1, & September\ 11,\ 2001 \leq t \leq December\ 31,\ 2001 \\ 0, & Otherwise \end{cases} \tag{2}$$

$$\mathscr{Z}_{Iraq,t} = \begin{cases} 1, & t \geq January\ 07,\ 2002 \\ 0, & Otherwise \end{cases} \tag{3}$$

respectively for terrorist attach to Twin Towers (September 11/2001) and for Iraqi attach (January 07/2002). Instead, the step function used for retail market is the following:

$$\mathscr{Z}_{ENI,t} = \begin{cases} 1, & t \geq October\ 06,\ 2004 \\ 0, & Otherwise \end{cases} \tag{4}$$

The results from the estimate of intervention models in the two markets are in Table 4. The error correction term is significant in both markets. The adjustment coefficients have the expected negative sign, which implies that they indeed reflect an error correction mechanism that tends to bring the system closer to its long-run equilibrium.

The two markets will now be examined separately.

[6] As described in Sect. 2, in order to select the optimal lag length we employ the Akaike Information Criterion (AIC) with Bayesan Schwartz (BIC) and Hannan-Quinn (HQC) in the VAR framework. Models include one lag.

Table 4 Parameter estimates for the full sample

Wholesale market		Retail market	
Input variable	Estimate	Input variable	Estimate
Constant	0,0096[c]	Constant	−0,0064[a]
EC	−0,0630[c]	EC	−0,1499[a]
$\Delta Brent$	0,0563	$\Delta Gaso$	0,1880[a]
$\Delta Platt$	0,0116	$\Delta Platt$	0,1079[a]
$\mathscr{Z}_{Terrorist}$	−0,0384[b]	$\mathscr{Z}_{Eni}$	0,0033[b]
$\mathscr{Z}_{Iraq}$	−0,0085		

[a] significant at 1%.
[b] significant at 5%.
[c] significant at 10%.

In the wholesale market the results appear to be in contrast. The short run effects of Platt's and oil prices turn out to be small and statistically insignificant. In fact, we confirm the presence of a long run equilibrium in the relationship between Platt's quotation and oil price but there is no short run relationship. It appears to be inelastic in the short run. Moreover, only one of the two exogenous shocks we have considered has a significant impact. This is the shock connected to the decline of oil prices following the September 11/2001 terrorist attacks on the United States ($\mathscr{Z}_{Terrorist,t}$). On the contrary, the military action in Iraq does not have a significant effect ($\mathscr{Z}_{Iraq,t}$). However in the retail market the coefficients estimated in the relationship between gasoline prices and Platt's quotations appear to be significant. This indicates that the gasoline price depends both from Platt's price and from itself. In fact, it appears to be autoregressive. Furthermore, the new price policy adopted by ENI (and followed by other companies) has had an increase in gasoline prices. In this market, as expected, it is impossible to observe the feedback effect between Platt's and gasoline prices. The estimated coefficients are omitted because they appear to be statistically insignificant.

4 Concluding Remarks

Our paper presents a preliminary result of an empirical analysis on the impact of political and/or corporate choices on the two stages of the gasoline production–distribution chain, using weekly data from January/2000 to November/2008. Considering some of the shocks singled out by EIA in the wholesale market and the new price policy adopted by ENI in the retail one, we have shown that using a simple VECM with intervention variables, the geopolitical events and industrial policies play a significant role in explaining the gasoline price adjustment. More specifically, shock in retail market contributes significantly to the increasing price of gasoline. It is generally assumed that consumers are more sensitive to price asymmetries[7] rather

[7] Many consumers, in fact, commonly complain that gasoline prices rise more quickly when crude oil prices are rising than they fall when crude oil prices are falling, exhibiting an asymmetric relationship.

than to the effects of exogenous shocks. In this paper we demonstrate that the price that consumers pay for fuel consumption is affected by the impact of events that cannot be controlled by oil companies. The effects of the new price policy adopted by the Italian market leader is quite relevant. In the retail market, in fact, ENI's price method has improved the average gasoline price that weighs on the consumers' budget. For this reason, we think that price asymmetries, that we have explored in other paper (Romano and Scandurra 2009), are not so relevant as the impact of policy choices in the gasoline markets. Future research efforts, which could eliminate some of the limitations of this study, could be (i) the impact of other shocks individuated by EIA; (ii) the effect of extreme volatility in gasoline price; (iii) the symmetry of shocks in the production–distribution chain.

References

Autorita' Garante della Concorrenza e del Mercato. (2007). *I prezzi dei carburanti in rete*. Retrieved Oct 25, 2009, from http://www.agcm.it/agcm_ita/news/news.nsf/4bdc4d49ebe1599dc12568da004b793b/6cc0ba483d4494d0c125726c005803bc/FILE/I681avvio.pdf.

Apergis, N., & Miller, S. M. (2009). Do structural oil-market shocks affect stock prices? *Energy Economics*, *31*, 569–575.

Engle, R. F., & Granger, C. W. (1987). Cointegration and error correction: Representation, estimation and testing. *Econometrica*, *55*, 251–276.

Gregory, A. W., & Hansen, B. E. (1996). Residual–Based tests for cointegration in models with regime shifts. *Journal of Econometrics*, *70*, 99–126.

Gregory, A. W., Nason, J. M., & Watt, D. G. (1996). Testing for structural breaks in cointegrated relationships. *Journal of Econometrics*, *71*, 321–341.

Hamilton, J. D. (1983). Oil and the macroeconomy since World War II. *The Journal of Political Economy*, *9*, 228–248.

Hamilton, J. D. (2003). What is an oil shock? *Journal of Econometrics*, *113*, 363–398.

Hamilton, J. D., & Herrera, M. A. (2002) Oil shocks and aggregate macroeconomic behaviour. *Journal of Money, Credit and Banking*, *35*, 265–286.

Hondroyiannis, G. (2004). Estimating residential demand for electricity in Greece. *Energy Economics*, *26*, 319–334.

Jones, C., & Kaul, G. (1996). Oil and the stock markets. *Journal of Finance*, *51*, 463–491.

Kim, D., & Perron, P. (2009). Unit root tests allowing for a break in the trend function at an unknown time under both the null and alternative hypotheses. *Journal of Econometrics*, *148*, 1–13.

Luktepol, H. (2005). *New introduction to multiple time series analysis*. Berlin: Springer.

Perron, P. (1989). The great crash, the oil price shock, and the unit-root hypothesis. *Econometrica*, *57*, 1361–1401.

Romano, A. A., & Scandurra, G. (2009). Price asymmetries in the Italian retail and wholesale gasoline markets, SIS Meeting "Statistical Methods for the analysis of large data-sets", Pescara, September 23–25.

Rodriguez, R. J. (2008). The impact of oil price shocks: Evidence from the industries of six OECD countries. *Energy Economics*, *30*, 3095–3108.

Zivot, E., & Andrews, D. W. K. (1992). Further evidence on the great crash, the oil price shock, and the unit-root hypothesis. *Journal of Business and Economic Statistics*, *10*, 25–43.

Part III
Nonparametric Kernel Estimation

Probabilistic Forecast for Northern New Zealand Seismic Process Based on a Forward Predictive Kernel Estimator

Giada Adelfio and Marcello Chiodi

Abstract In seismology predictive properties of the estimated intensity function are often pursued. For this purpose, we propose an estimation procedure in time, longitude, latitude and depth domains, based on the subsequent increments of likelihood obtained adding an observation one at a time. On the basis of this estimation approach a forecast of earthquakes of a given area of Northern New Zealand is provided, assuming that future earthquakes activity may be based on the smoothing of past earthquakes.

1 Introduction

Earthquakes forecast is defined as a vector of earthquakes rates corresponding to specified multi-dimensional intervals (defined by the location, time and magnitude) (Geller 1997). From the rates specified in each forecast a vector of expected number of events within the time interval for all intervals is calculated. The expected number is just the earthquake rate multiplied by the volume in parameter space of the bin; the expectations are dimensionless and they correspond directly to earthquake rates per unit area, magnitude and time.

Expectations is sometimes referred to as prediction, but in earthquakes contexts it has different meaning, since earthquakes prediction implies high probability and imminence. In few words earthquake prediction is considered as a special case of forecast in which the forecast rate is temporarily high enough to justify an exceptional response beyond that appropriate for normal conditions (Geller 1997).

In this paper we provide a forecast for magnitude 4.5 and larger earthquakes of Northern New Zealand. The forecast is retrospective (i.e., after that events occurred) and uses data available on a fixed inception date. This approach should be considered as a statistical forecast in the sense that it estimates the probability of occurrence of further events given the past space-time history per unit area and time; as suggested in Kagan and Jackson (2000), it may be useful just for scientific testing and therefore, it could not be used for official warnings.

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_14,

Here we assume that the forecast is proportional to the intensity function of a four dimensional process, obtained by smoothing data in space (3D) and time domains. The intensity function of the process is estimated by a variation of kernel estimators approach (Silverman 1986), with good properties in terms of *MISE* and valid predictive features. The proposed approach (Forward Likelihood-based Predictive (FLP) approach) for nonparametric estimation in space-time point process, can be considered as a generalization of cross-validation, but aimed to prediction accounting for the temporal ordering of observations.

The forecast is provided for events from 1994 to 2008 on the basis of a smoothing of earthquakes occurred from 1951 to 1994 in the same area and the same magnitude threshold.

Main features of point processes are reminded in Sect. 2. The flexible estimation procedure used in this paper is introduced in Sect. 3 and forecast results for Northern New Zealand are showed in Sect. 4. In Sect. 5 some conclusions and direction for future work are provided.

2 Conditional Intensity Function of Point Processes

Because of random feature of earthquakes catalogs, point process theory is often used. Indeed each earthquake is identified by a point in space (hypocentral coordinates), in time (occurrence time) and magnitude domain. Basic models to describe the seismic rate of a given region, assume that earthquakes occur in space and time according to a stationary point process, i.e., Poisson process such that the conditional rate is constant. Of course the stationarity hypothesis may be acceptable only in time and for main events, since hypocenters are usually spatially heterogeneous and clustered; moreover aftershocks and foreshocks (i.e., event that might occur after and before a main event respectively) are also an evident proof of the dependence of seismic activity on the past. Therefore the introduction of more complex models than stationary Poisson process is often necessary, relaxing the assumption of statistical independence of earthquakes.

Point processes can be specified mathematically in several ways, for instance, by considering the joint distributions of the counts of points in arbitrary sets or defining a complete intensity function, that is a function of the points history that generalizes the rate function of a Poisson process.

More formally a space-time point process is a random collection of points, each representing location and time of a single event.

Let N be a point process on a spatial-temporal domain $X \subseteq \mathbb{R}^d \times \mathbb{R}_+$ and e(x) be the Lebesgue measure of x; its conditional intensity function is defined by:

$$\lambda(t, \mathbf{s}|\mathscr{H}_t) = \lim_{\ell(\delta t)\ell(\delta \mathbf{s})\to 0} \frac{E[N([t, t+\delta t) \times [\mathbf{s}, \mathbf{s}+\delta \mathbf{s})|\mathscr{H}_t)]}{\ell(\delta t)\ell(\delta \mathbf{s})} \tag{1}$$

where $\mathscr{H}_t$ is the space-time occurrence history of the process up to time t, i.e., the σ-algebra of events occurring at times up to but not including t; $\delta t, \delta \mathbf{s}$ are time and space increments respectively, and $E[N([t, t+\delta t) \times [\mathbf{s}, \mathbf{s}+\delta \mathbf{s})|\mathscr{H}_t)]$ is the history-dependent expected value of occurrence in the volume $\{[t, t+\delta t) \times [\mathbf{s}, \mathbf{s}+\delta \mathbf{s})\}$.

The conditional intensity function uniquely characterizes the finite-dimensional distributions of point processes (Daley and Vere-Jones 2003). For instance, self-exciting point processes are used to model events that are clustered together; self-correcting processes, e.g., the stress-release model (Vere-Jones 1978), are suggested when regularities are observed.

To assess the goodness of fit diagnostic tests could be used (see Adelfio and Schoenberg 2009; Adelfio and Chiodi 2009).

3 Kernel-Based Models

To provide a valid forecast of the seismic activity of a fixed area we need the definition of a valid stochastic model.

Parametric estimation could not be always useful, since it requires the definition of a reliable mathematical model from the geophysical theory; for this reason, in this paper some techniques for estimating the intensity function of space-time point processes are developed, based on generalizations of (anisotropic and isotropic) kernel estimators.

Indeed some disadvantages of the parametric modelling can be overcome by a flexible procedure (nonparametric technique), based on kernel density methods (Silverman 1986). Given n observed events $\mathbf{s}_1, \mathbf{s}_1 \ldots, \mathbf{s}_n$ in a d-dimensional given region, the kernel estimator of the unknown density f in $\mathbb{R}^d$ is defined as:

$$\hat{f}(s_1, \ldots, s_d; \mathbf{h}) = \frac{1}{n h_{s_1} \cdots h_{s_d}} \sum_{i=1}^{n} K\left(\frac{s_1 - s_{i1}}{h_{s_1}}, \cdots, \frac{s_d - s_{id}}{h_{s_d}}\right) \tag{2}$$

where $K(s_1, \ldots, s_d)$ denotes a multivariate kernel function operating on d arguments centered at $(s_{i1}, \ldots, s_{id})$ and $(h_{s_1}, \ldots, h_{s_d})$ are the smoothing parameters of kernel functions. If $\mathbf{s}_i = \{t_i, x_i, y_i, z_i\}$, the space-time kernel intensity estimator is defined by the superposition of the separable kernel densities:

$$\hat{\lambda}(t, x, y, z; \mathbf{h}) \propto \sum_{i=1}^{n} K_t\left(\frac{t - t_i}{h_t}\right) K_s\left(\frac{x - x_i}{h_x}, \frac{y - y_i}{h_y}, \frac{z - z_i}{h_z}\right) \tag{3}$$

where K_t and K_s are temporal and spatial kernel density functions, as in (2), respectively. Hence the advantage of this approach is to make simpler the complex issue of estimating the conditional intensity function of a particular point process described by (1), dealing just with the intensity function of a simple inhomogeneous Poisson process identified by a space-time Gaussian kernel intensity like in (3) as in Adelfio and Ogata (2010), Adelfio (2010a) and Adelfio (2010b).

In nonparametric approaches based on kernel estimators the estimation of the smoothing parameter $\mathbf{h}$ is the crucial point. In this paper we propose an estimation method that is a general version of cross-validation techniques. In Adelfio et al. 2006 the seismicity of the Southern Tyrrhenian Sea is described by the use of Gaussian kernels and the optimum value of $\mathbf{h}$ is chosen such as to minimize the mean integrated square error (MISE) of the estimator $\hat{f}(\cdot)$. In particular the authors used the value $\mathbf{h}_{opt}$ that in Silverman (1986) is obtained minimizing the MISE of $\hat{f}(\cdot)$ assuming multivariate normality. In Adelfio and Ogata 2010 a naive likelihood cross-validation is optimized to obtain the bandwidth of the smoothing kernel used to estimate the intensity for earthquake occurrence of Northern Japan. In Adelfio 2010a the same area is analyzed by smoothing data according to a variable bandwidth procedure: the bandwidth for the jth event, $j = 1, \ldots, n$ is $h_j = (h_x^j, h_y^j, h_t^j)$ that is the radius of the smallest circle centered at the location of the jth event (x_j, y_j, t_j) that includes at least a fixed number np of other events.

3.1 Forward Likelihood for Prediction (FLP) Estimation

In this paper we propose a space-time estimation procedure aimed both at the fit of past data and the prediction of future ones, based on increments of log-likelihood adding single event at a time.

The approach accounts for ordering in time of the observed process and deals with predictive properties of estimates.

More formally, let $\mathbf{s}_i$ be m points observed in the space region $\Omega_\mathbf{s}$ and in the period of time $(T_0 - T_{max})$ with $\mathbf{h}$ the vector of smoothing parameters. The log-likelihood of the point process observed up to time t_m in the observed space region $\Omega_\mathbf{s}$ is:

$$\log L(\hat{\mathbf{h}}(H_{t_m}); I_{t_m}) = \sum_{i=1}^{m} \log \hat{\lambda}(\mathbf{s}_i; \hat{\mathbf{h}}(H_{t_m})) - \int_{T_0}^{t_m} \int_{\Omega_\mathbf{s}} \hat{\lambda}(\mathbf{s}; \hat{\mathbf{h}}(H_{t_m})) d\mathbf{s} dt \quad (4)$$

where $\hat{\lambda}(\cdot)$ is defined in (3) and depends on the unknown parameters estimated by $\hat{\mathbf{h}}(H_{t_m}) = \hat{\mathbf{h}}(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m)$ and I_{t_k} is the time region from T_0 to t_k. On the basis of the defined notation, the likelihood computed on the first $m + 1$ observations but using the estimates based on first m observations is defined as:

$$\log L(\hat{\mathbf{h}}(H_{t_m}); I_{t_m}) = \sum_{i=1}^{m} \log \hat{\lambda}(\mathbf{s}_i; \hat{\mathbf{h}}(H_{t_m})) - \int_{T_0}^{t_m} \int_{\Omega_\mathbf{s}} \hat{\lambda}(\mathbf{s}; \hat{\mathbf{h}}(H_{t_m})) d\mathbf{s} dt \quad (5)$$

Then, we use the difference between (4) and (5) to measure the *predictive information* of the first m observations on the $(m + 1)$-th:

$$\delta_{m,m+1} \equiv \log L(\widehat{\mathbf{h}}(H_{t_m}); I_{t_{m+1}}) - \log L(\widehat{\mathbf{h}}(H_{t_m}); I_{t_m}).$$

Therefore, we choose $\widetilde{\mathbf{h}}(H_{t_m})$ which maximizes

$$FLP_{m_1,m_2}(\hat{\mathbf{h}}) = \sum_{m=m_1}^{m_2} \delta_l(\hat{\mathbf{h}}(H_{t_m}); H_{t_{m+1}}):$$

$$FLP_{m_1,m_2}(\tilde{\mathbf{h}}) \geq FLP_{m_1,m_2}(\hat{\mathbf{h}}) \quad \forall \hat{\mathbf{h}} \in \Theta \tag{6}$$

with $m_2 = m - 1$ and m_1 is such that $t_{m_1} - T_0 \approx \frac{T_{max}-T_0}{2}$.

Although the method could be considered as a direct extension of the idea of prediction, it is almost heuristic. In first applications we assumed constant vector of parameters for each point observed in a two dimensional space and time (Adelfio and Chiodi 2011), with good results both in terms of likelihood and standard deviation. Here we use the proposed technique to estimate the parameters of a four dimensional kernel intensity in the longitude-latitude-depth-time domain:

$$\hat{f}(\mathbf{s}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{|\mathbf{h}|}(2\pi)^{-d/2} \exp\left\{-\frac{1}{2}(\mathbf{s}-\mathbf{s}_i)^T(\mathbf{h})^{-1}(\mathbf{s}-\mathbf{s}_i)\right\}$$

where $d = 4$, $\mathbf{s} = \{x, y, z, t\}$ and $\mathbf{h}$ is a matrix of smoothing parameters for anisotropic space-time kernel, such that:

$$\mathbf{h} = \begin{pmatrix} h_x & h_{xy} & h_{xz} & 0 \\ h_{xy} & h_y & h_{yz} & 0 \\ h_{xz} & h_{yz} & h_z & 0 \\ 0 & 0 & 0 & h_t \end{pmatrix}$$

4 FLP-model and Forecast for Northern New Zealand

To provide a forecast of an active seismic area, we selected a subset of the GeoNet catalog of New Zealand earthquakes. Completeness issues of this catalog are discussed in Harte and Vere-Jones (1999). The data consist of earthquakes of magnitude 4.5 and larger that are chosen from the wide region $-41.5° \sim -37°$N and $174.5° \sim 176.5°$E and for the time span 1951–2008. That area is characterized by several deep events, with depth down to 530 km and is of great interest for geologists because of the high rate of activity recorded in years.

The forecasts for earthquakes occurred in 1994–2008 is based on a smoothing of past activity (1951–1994).

In Figs. 1–3, the smoothed intensity function and forecasts for activity in 1994–2008 and increasing level of depth are showed. Indeed it could be interesting to check how this variable influence results, since the behavior of deep events is often not well understood, because of the complex focal mechanism at high level of depth.

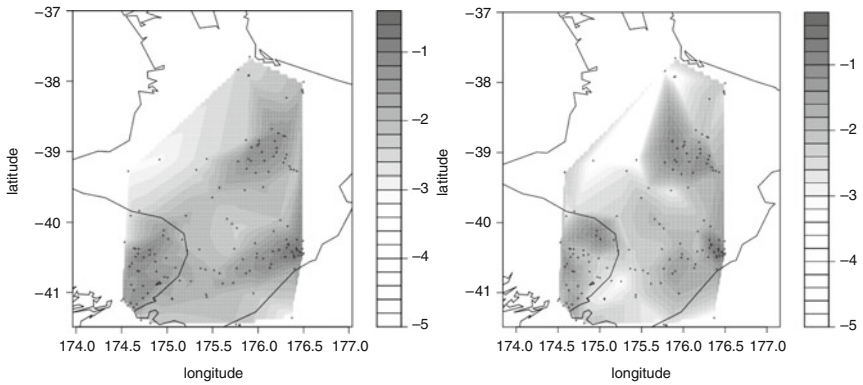


Fig. 1 Smoothed intensity function (log scale) for the second set of events (1994–2008; *on the left*) and forecast (*on the right*) for depths up to 100.00 km

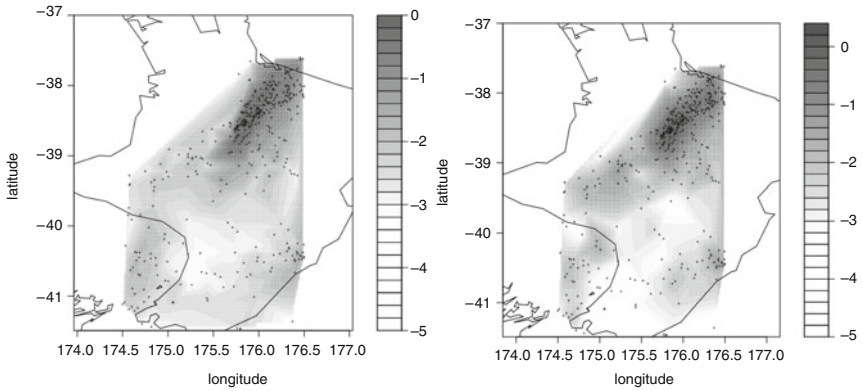


Fig. 2 Smoothed intensity function (log scale) for the second set of events (1994–2008; *on the left*) and forecast (*on the right*) for depths up to 200.00 km

From these figures, a migration of activity from one part of the analyzed region to another, and between depths, is observed. Moreover, from the correspondence of the provided images, forecast results for the time interval 1994–2008 seem to match with real observed occurrence of events in the same time. Indeed the provided forecasts identify main areas of occurrence and provide a very realistic description of observed seismicity in the considered time interval. Statistical tests to asses the validity of the forecasts are provided in Table 1, comparing forecast results between FLP and Silverman's rule based on χ^2 test.

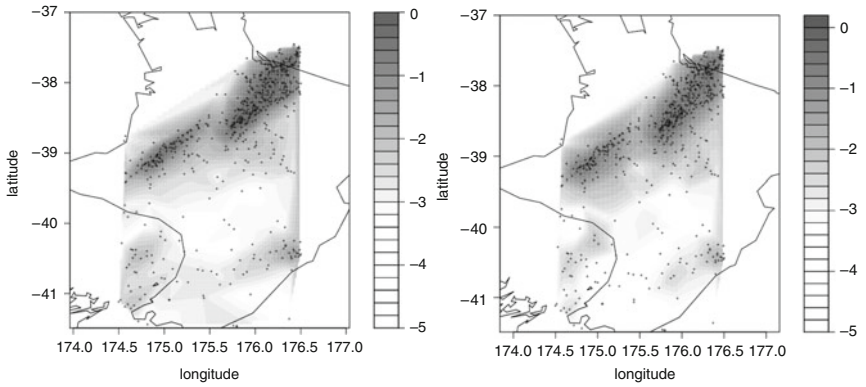


Fig. 3 Smoothed intensity function (log scale) for the second set of events (1994–2008; *on the left*) and forecast (*on the right*) for depths up to 600.00 km

Table 1 Comparative forecast results between FLP and Silverman's rule based on χ^2 test for a two dimensional grid (anisotropic in both cases), using the Berman-Turner quadrature approximation (Baddeley and Turner 2000) that requires a quadrature scheme consisting of the original data point pattern, an additional pattern of dummy points, and a vector of quadrature weights for all these points. Similar results are observed for the space-time case (3D)

Depth	n	Method	X^2	p-value
$z \leq 100$ km	144	FLP	5.826	0.771
		Silverman	33.608	0.002
$z \leq 200$ km	394	FLP	11.191	0.670
		Silverman	52.302	0.000
$z \leq 600$ km	577	FLP	15.037	0.375
		Silverman	63.217	0.000

5 Comments and Conclusion

In this paper, we introduced an estimation procedure based on increments of likelihood, taking into account the contribution to the log-likelihood function of forward event, given the nonparametric estimates based on the previous ones.

From observed results the provided nonparametric approach makes possible a reasonable characterization of seismicity, since it does not constrain the process to have predetermined properties. Indeed, the estimated model seems to follow adequately the seismic activity of the observed areas, characterized by highly variable changes both in space and in time and because of its flexibility, it provides a good fitting to local space-time changes as just suggested by data.

Moreover, it seems that the used model provides valid forecast results for the studied area, although these results can be considered as just an empirical and first try for this complex issue. Indeed, the proposed forecast approach is not based on geophysical or geological basis, but it just assumes that further events occurrence is proportional to a smoothing of previous events.

In other words, flexible kernel estimators are here used to estimate a four dimensional intensity function, without specifying any forecast models.

As a direction for future work, we think that methods to test the reliability of these forecast should be developed, as in Jackson and Kagan (1999).

References

Adelfio, G. (2010a). An analysis of earthquakes clustering based on a second-order diagnostic approach. In Palumbo, et al. (Eds.), *Data analysis and classification* (pp. 309–317). Series: Studies in classification, data analysis, and knowledge organization. Springer Berlin Heidelberg.

Adelfio, G. (2010b). Kernel estimation and display of a five-dimensional conditional intensity function. *Nonlinear Processes Geophysics*, *17*, 1–8.

Adelfio, G., & Chiodi, M. (2009). Second-order diagnostics for space-time point processes with application to seismic events. *Environmetrics*, *20*, 895–911.

Adelfio, G., & Chiodi, M. (2011). Kernel intensity for space-time point processes with application to seismological problems. In Fichet, et al. (Eds.), *Classification and multivariate analysis for complex data structures* (pp. 401–408). Series: Studies in classification, data analysis, and knowledge organization. Springer Berlin Heidelberg, ISBN: 978-3-642-13312-1.

Adelfio, G., Chiodi, M., De Luca, L., Luzio, D., & Vitale, M. (2006). Southern-tyrrhenian seismicity in space-time-magnitude domain. *Annals of Geophysics*, *49*(6), 1245–1257.

Adelfio, G., & Ogata, Y. (2010). Hybrid kernel estimates of space-time earthquake occurrence rates using the ETAS model. *Annals of the Institute of Statistical Mathematics*, *62*(1), 127–143.

Adelfio, G., & Schoenberg, F. P. (2009). Point process diagnostics based on weighted second-order statistics and their asymptotic properties. *Annals of the Institute of Statistical Mathematics*, *61*(4), 929–948.

Baddeley, A., & Turner, T. R. (2000). Pratical maximum pseudo likelihood for spatial point patterns (with discussion). *Australian & New Zealand Journal of Statistics*, *42*(3), 283–322.

Daley, D. J., & Vere-Jones, D. (2003). *An introduction to the theory of point processes* (2nd edn.). New York: Springer-Verlag.

Geller, R. J. (1997). Earthquake prediction: a critical review. *Geophysical Journal International*, *131*(3), 425–450.

Harte, D., & Vere-Jones, D. (1999). Differences in coverage between the pde and new zealand local earthquake catalogues. *New Zealand Journal of Geology and Geophysics*, *42*, 237–253.

Jackson, D., & Kagan, Y. (1999). Testable earthquake forecasts for 1999. *Seismological Research Letters*, *80*, 393–403.

Kagan, Y., & Jackson, D. (2000). Probabilistic forecasting of earthquakes. *Geophysics Journal International*, *143*, 438–453.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Vere-Jones, D. (1978). Earthquake prediction - a statistician's view. *Journal of Physics Earth*, *26*, 129–146.

Discrete Beta Kernel Graduation of Age-Specific Demographic Indicators

Angelo Mazza and Antonio Punzo

Abstract Several approaches have been proposed in literature for the kernel graduation of age-specific demographic indicators. Nevertheless, although age is pragmatically a discretized variable with a finite support (typically age at last birthday is considered), commonly used methods employ continuous kernel functions. Moreover, symmetric kernels, that bring in further bias at the support boundaries (the so-called problem of boundary bias), are routinely adopted. In this paper we propose a discrete kernel smooth estimator specifically conceived for the graduation of discrete finite functions, such are age-specific indicators. Kernel functions are chosen from a family of conveniently discretized and re-parameterized beta densities; since their support matches the age range, the issue of boundary bias is eliminated. An application to 1999–2001 mortality data from the Valencia Region (Spain) is also presented.

1 Introduction

Many demographic phenomena such as mortality, fertility, nuptiality and migration are strongly related to age, in the sense that the intensity of events varies sharply across the age range (Preston et al. 2001). Therefore, a key aspect of demographic research consists in studying such age-specific patterns in order to discover regularities and make comparisons across time and space. Although in the following we will only refer to mortality, our approach can easily be extended to the other age-dependent phenomena.

Let X be the variable age, with finite support $\mathscr{X} = \{0, 1, \ldots, \omega\}$, being ω the maximum age of death. We consider X as discrete, although age is in principle a continuous variable, since in demographic and actuarial applications age at last birthday is generally used; furthermore, discrete are also the commonly used age-specific indicators, such as the *proportion dying* q_x which is the proportion of persons who die at age x, $x \in \mathscr{X}$, and the *number dying* d_x which is the number of persons that would die at age x if starting from an arbitrary large hypothetical cohort l_0 (conventionally $l_0 = 100{,}000$ is used).

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_15,

The use of such age-specific indicators is often deemed as not appropriate since a specific observed pattern may be often intended as a single realization of a stochastic phenomenon with certain distinctive mortality traits. Such random fluctuations are more of concern in actuarial studies and in applied demography, when small area datasets are investigated or when the events under investigation are particularly rare. To cope with this issue, several graduation techniques have been proposed in literature (see, e.g., Debòn et al. 2005 and Debòn et al. 2006 for an exhaustive comparison of parametric and nonparametric methods, respectively, in the graduation of mortality data), based on the assumption that if the number of individuals in the group on whose experience data are based were considerably larger, then the set of observed indicators would display a much more regular progression with age (Copas and Haberman 1983). In this context, the discretization of age could also come handy to actuaries that, in calculating life insurance premiums, annuities, reserves, and so on, have to produce "discrete" graduated mortality tables starting from the observed counterparts.

The most popular statistical method for nonparametric graduation is the kernel smoothing (Copas and Haberman 1983); in such method, most of the attention is usually dedicated to the selection of the smoothing parameter while symmetric kernel functions are routinely used. Nevertheless, if the use of symmetric kernels is appropriate when fitting functions with unbounded supports from both sides, its use is not adequate with age-dependent functions (Chen 1999, 2000). When smoothing is made near the boundaries, in fact, fixed symmetric kernels do allocate weight outside the support (e.g., negative or unrealistic high ages) causing the well-known problem of boundary bias.

In this paper, in Sect. 2, we propose a discrete kernel smooth estimator that we have specifically conceived for the graduation of discrete finite functions. This approach, by construction, overcomes the problem of boundary bias, given that the kernels are chosen from a family of conveniently discretized and reparameterized beta densities whose support $\mathscr{X}$ matches the age range. Finally, in Sect. 3, we present an application of our kernel estimator to 1999–2001 mortality data from the Valencia Region (Spain).

2 A Discrete Beta Kernel Estimator

Given the values y_x of the age-specific indicators, the following general form of a discrete kernel estimator is considered

$$\widehat{y}_x = \sum_{j=0}^{\omega} y_j k_h\left(j; m = x\right), \quad x \in \mathscr{X}, \tag{1}$$

where $k_h\left(\cdot; m\right)$ is the *discrete kernel function* (simply said *kernel*), m is the single mode of the kernel, and h is the so-called *smoothing parameter* governing the bias-variance trade-off. The quantities y_x, $x \in \mathscr{X}$, can be any discrete finite function

on $\mathscr{X}$, such as d_x or q_x. It is worth to note that while the graduation of d_x is, in principle, a problem of distribution estimation, the graduation of q_x is a problem of regression estimation; in these terms, the model (1) is "general" because it can handle both problems. Also it can be noted that, since we are treating age as being discrete, kernel graduation by means of (1) is equivalent to moving (or local) weighted average graduation (Gavin et al. 1995).

Model (1) can be conveniently written in the following compact form

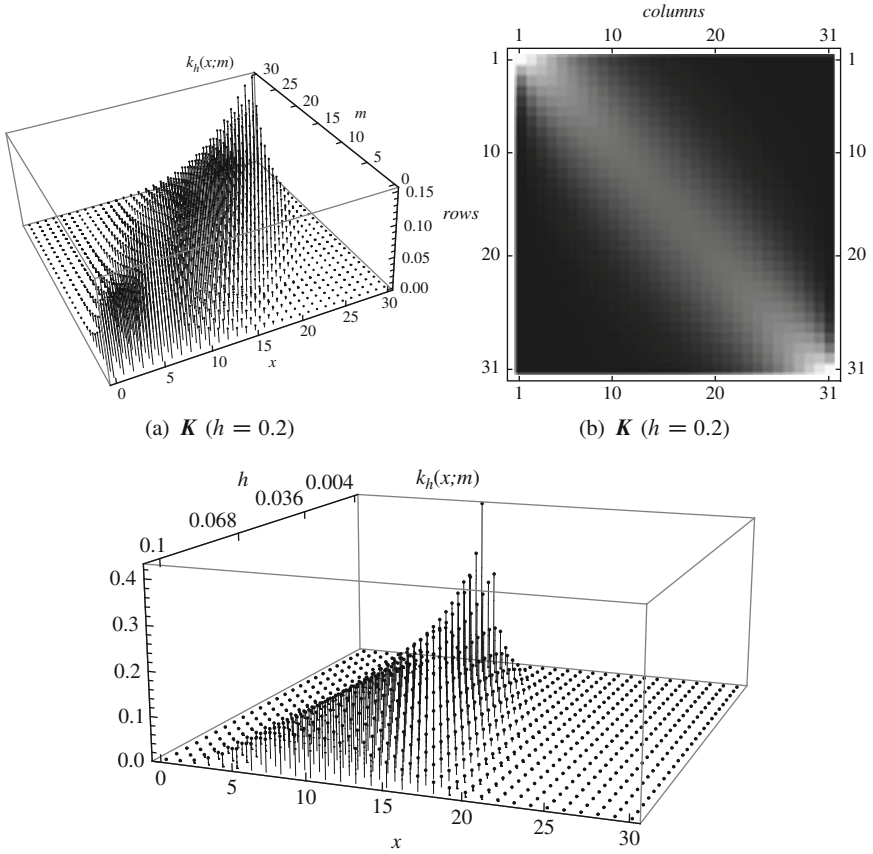
$$\widehat{\boldsymbol{y}} = \boldsymbol{K}\,\boldsymbol{y},$$

where $\boldsymbol{y}$ and $\widehat{\boldsymbol{y}}$ are, respectively, the $(\omega+1)$-dimensional vectors of "observed" and graduated indicators, while $\boldsymbol{K}$ is the so-called $(\omega+1)\times(\omega+1)$ *smoother* matrix (or *hat* matrix) in which the i-th row contains the $\omega+1$ weights allocated to the age-specific indicators y_x, $x \in \mathscr{X}$, in order to obtain $\widehat{y}_{i-1}$. Without loss of generality, we have constrained the age set of graduated specific indicators to be $\mathscr{X}$.

As system of weights in (1), we have chosen to adopt the following discrete beta distribution

$$k_h(x;m) = \frac{\left(x+\frac{1}{2}\right)^{\frac{m+\frac{1}{2}}{h(\omega+1)}}\left(\omega+\frac{1}{2}-x\right)^{\frac{\omega+\frac{1}{2}-m}{h(\omega+1)}}}{\sum_{j=0}^{\omega}\left(j+\frac{1}{2}\right)^{\frac{m+\frac{1}{2}}{h(\omega+1)}}\left(\omega+\frac{1}{2}-j\right)^{\frac{\omega+\frac{1}{2}-m}{h(\omega+1)}}}, \quad x \in \mathscr{X}, \tag{2}$$

with $m \in \mathscr{X}$ and $h > 0$. This distribution, introduced in Punzo and Zini (2011), is a discretization on $\mathscr{X}$ of a beta density defined on $[-1/2, \omega+1/2]$, conveniently re-parameterized, as in Punzo (2010), with respect to the mode m and another parameter h that is closely related to the distribution variability. From a nonparametric point of view, it is important to note that the probability mass function in (2) is smoothly non-increasing as $|x-m|$ increases. Keeping constant ω and h, Fig. 1a shows the discrete beta weights in the smoother matrix and, consequently, it illustrates the effect of varying the mode m; in order to give a visual representation of the values in $\boldsymbol{K}$, its matrix plot is also displayed in Fig. 1b. As regards the other parameter h, we have that, for $h \to 0^+$, $k_h(x;m)$ tends to a Dirac delta function in $x = m$, while for $h \to \infty$, $k_h(x;m)$ tends to a discrete uniform distribution; Fig. 1c shows the effect of varying h, maintaining constant ω and m. Thus h can be considered as the smoothing parameter of the estimator (1); indeed, as h becomes smaller, the spurious fine structure becomes visible, while as h gets larger, more details are obscured. With reference to the problem of distribution estimation, in order to obtain values of d_x summing to $\sum_{x\in\mathscr{X}} d_x$, we can estimate, in a first step, all the $\omega+1$ values of $\widehat{d}_x$ according to (1) and then, in a second step, we can normalize these as follows



(a) $\boldsymbol{K}$ ($h = 0.2$) (b) $\boldsymbol{K}$ ($h = 0.2$)

(c) Effect of h ($m = 10$)

Fig. 1 Discrete beta weights are shown in (**a**), matrix plot of $\boldsymbol{K}$ is displayed in (**b**) using different *gray-levels*, and the effect of h is illustrated in (**c**). In all plots $\omega = 30$

$$\widehat{d}_x = \frac{\widehat{d}_x}{\sum_{j=0}^{\omega} \widehat{d}_j} \sum_{j=0}^{\omega} d_j. \tag{3}$$

Discrete beta kernels possess two peculiar characteristics. Firstly, their shape, fixed h, automatically changes according to the value of m (see Fig. 1a). However, as it can be seen in Fig. 1a, two kernels having modes in m and $\omega - m$ are each other's reflection in $x = \omega/2$, that is, $k_h(x;m) = k_h(\omega - x; \omega - m)$, $x \in \mathscr{X}$. Secondly, the support of the kernels matches the age range $\mathscr{X}$ so that no weight is assigned outside the data support; this means that the order of magnitude of the bias does not increase near the boundaries and the so-called problem of boundary bias is automatically overcome.

Finally note that although the model (1), for the case of distribution estimation, is equal in philosophy to the nonparametric discrete kernel estimator proposed

in Punzo (2010), there is a substantial difference in the way the discrete beta kernels are adopted. In detail:

- in Punzo (2010) the nonparametric estimator is "globally" defined as a mixture, with observed weights y_m, of $\omega+1$ discrete beta components $k_h(x;m)$, of equal parameter h, with mode in m, $m \in \mathscr{X}$;
- in (1) the estimator of y_x is "locally" defined by placing a discrete kernel distribution at the point x for which we wish to estimate the true proportion and then forming a weighted average over all the age-specific proportions, where the weight attached to each age-specific proportion is the value of the discrete kernel distribution, with mode in x, at that age.

2.1 The Choice of the Smoothing Parameter

The literature on data-driven methods for selecting the optimal value for h is vast; however, cross-validation (Stone 1974) is without doubt the most commonly used and the simplest to understand. Cross-validation simultaneously fits and smooths the data by removing one data point at a time, estimating the value of the function at the missing point, and then comparing the estimate to the omitted, observed value. For a complete description of cross-validation in the context of graduation, see Gavin et al. (1995). The cross-validation statistic or score, $CV(h)$, for model (1) is

$$CV(h) = \frac{1}{n}\left(\boldsymbol{y} - \widehat{\boldsymbol{y}}^{(-x)}\right)'\left(\boldsymbol{y} - \widehat{\boldsymbol{y}}^{(-x)}\right) = \frac{1}{n}\sum_{x\in\mathscr{X}}\left(y_x - \widehat{y}_x^{(-x)}\right)^2, \qquad (4)$$

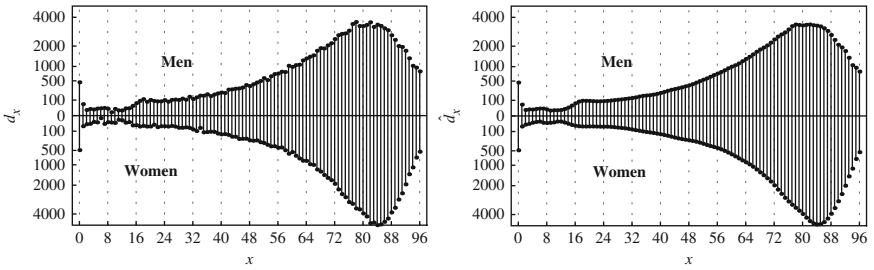
where

$$\widehat{y}_x^{(-x)} = \frac{\sum\limits_{\substack{j\in\mathscr{X}\\ j\neq x}} y_j k_h(j;m=x)}{1-k_h(x;m=x)} = \frac{\sum\limits_{\substack{j\in\mathscr{X}\\ j\neq x}} y_j k_h(j;m=x)}{\sum\limits_{\substack{j\in\mathscr{X}\\ j\neq x}} k_h(j;m=x)}$$

is the estimated value at age x computed by removing the age-specific indicator y_x at that age. The value of h that minimizes $CV(h)$ is referred to as the cross-validation smoothing parameter, $\widehat{h}_{CV}$.

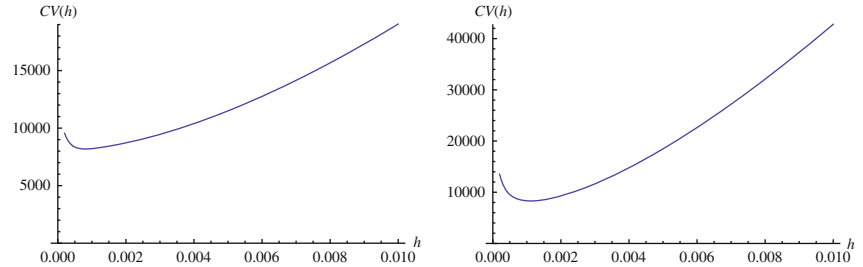
3 An Application to Mortality Data

The model described in the previous section is applied to mortality data from the Valencia Region for the period 1999–2001. Data are classified by age (ranging from 0 to 100 or older) and sex (source: Spanish National Institute of Statistics, INE). In Debòn et al. (2005) and Debòn et al. (2006) the authors use the same dataset in order to compare parametric and nonparametric methods for the graduation of



(a) "Observed" (b) Graduated

Fig. 2 Observed and graduated men-to-women rootplots (barplots plotted on a square root scale) of the number dying per 100,000 hypothetical live births



(a) Men: $\hat{h}_{CV} = 0.000802$ (b) Women: $\hat{h}_{CV} = 0.001093$

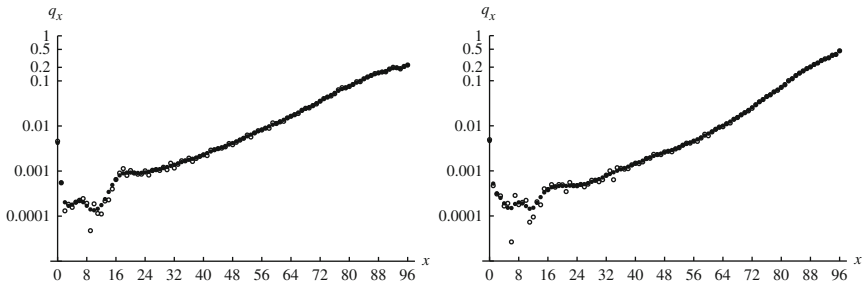
Fig. 3 Plot of $CV(h)$ as a function of h

mortality rates q_x; in analogy with them, we have chosen to take a range of ages between 0 and $\omega = 96$. In the period and range under study, there were nearly 3.96 million men at risk and 4.11 million women. Other summary information can be gained in Debòn et al. (2005) and Debòn et al. (2006).

In Fig. 2a, the observed men-to-women rootplots (barplots plotted on a square root scale) are depicted. Rootplots for either sexes are slightly ragged and negatively skewed, with a long tail over lower ages; infant mortality, at $x = 0$, is also distinguishable. Above all with reference to the male population, a small but prominent bump over the age of 18 is also visible; this "excess mortality" is probably due to an increase in a variety of risky activities, the most notable being to obtain a driver's license.

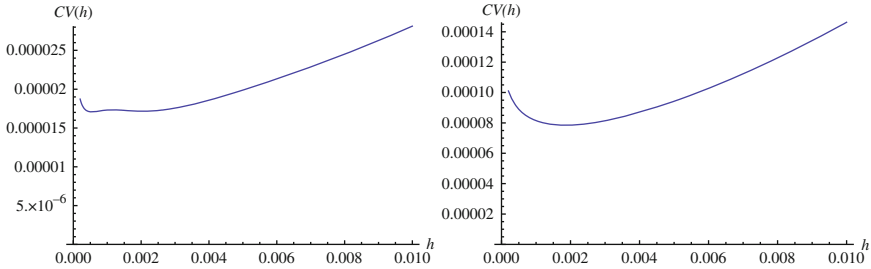
All of these characteristics are preserved and emphasized in Fig. 2b where the discrete normalized beta kernel estimator (3) is fitted, both for men and women, to the observed data. Smoothing parameter h for men, and women, was estimated by minimizing in `Mathematica` the quantity (4). The cross-validation estimated values of h, as well as the graphical representation of $CV(h)$ plotted against h, are displayed in Fig. 3.

The same general considerations hold also for the analysis of the mortality rates. Fig. 4 shows, in logarithm scale, the proportion dying for male and female



(a) Men (b) Women

Fig. 4 Observed (∘) and graduated (•) proportion dying in logarithm scale



(a) Men: $\hat{h}_{CV} = 0.002046$ (b) Women: $\hat{h}_{CV} = 0.001842$

Fig. 5 Plot of $CV\,(h)$ as a function of h

population, and superimposes them the graduated counterparts. It is easy to note that the graduated points have a more regular behavior than the observed ones for the age range between 0 and 40. As in the previous case, the two different values for h were obtained minimizing in `Mathematica` the cross-validation statistics (4). The cross-validation estimated values of h, as well as the graphical representation of $CV\,(h)$ plotted against h, are displayed in Fig. 5.

4 Concluding Remarks

Demographic phenomena usually exhibit a rather complex age pattern, and statistical models used in graduation should be quite flexible in order to capture such complexity. Moreover, data usually available refer to age at last birthday, that is a discretized variable with a finite support. In this paper we have proposed a discrete kernel estimator specifically conceived for the smoothing of discrete finite functions. Kernel functions were chosen from a family of conveniently discretized and reparameterized beta densities proposed in Punzo (2010); since their support matches the age range boundaries, the estimates are free of boundary bias. Moreover, the resulting discrete beta kernel graduation method is conceptually simple and so is its implementation.

References

Chen, S. X. (1999). Beta kernel estimators for density functions. *Computational Statistics and Data Analysis*, *31*(2), 131–145.

Chen, S. X. (2000). Beta kernel smoothers for regression curves. *Statistica Sinica*, *10*(1), 73–91.

Copas, J. B., & Haberman, S. (1983). Non-parametric graduation using kernel methods. *Journal of the Institute of Actuaries*, *110*, 135–156.

Debòn, A., Montes, F., & Sala, R. (2005). A comparison of parametric models for mortality graduation. Application to mortality data for the Valencia region (Spain). *Statistics and Operations Research Transactions*, *29*(2), 269–288.

Debòn, A., Montes, F., & Sala, R. (2006). A comparison of nonparametric methods in the graduation of mortality: Application to data from the Valencia region (Spain). *International Statistical Review*, *74*(2), 215–233.

Gavin, J. B., Haberman, S., & Verrall, R. J. (1995). Graduation by kernel and adaptive kernel methods with a boundary correction. *Transactions of the Society of Actuaries*, *47*, 173–209.

Preston, S., Heuveline, P., & Guillot, M. (2001). *Demography: Measuring and modelling population processes*. Oxford: Blackwell.

Punzo, A. (2010). Discrete beta-type models. In H. Locarek-Junge & C. Weihs (Eds.), *Studies in Classification, Data Analysis, and Knowledge Organization: Classification as a Tool for Research, Part 2*, (pp. 253–261). Berlin: Springer.

Punzo, A., & Zini, A. (2011). Discrete approximations of continuous and mixed measures on a compact interval. *Statistical Papers*, (to appear).

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B*, *36*(1), 111–147.

Kernel-Type Smoothing Methods of Adjusting for Unit Nonresponse in Presence of Multiple and Different Type Covariates

Emilia Rocco

Abstract This paper deals with the nonresponse problem in the estimation of the mean of a finite population following a nonparametric approach. Weighting adjustment is a popular method for handling unit nonresponse. It operates by increasing the sampling weights of the respondents in the sample using estimates of their respond probabilities. Typically, these estimates are obtained by fitting parametric models relating response occurrences and auxiliary variables. An alternative solution is the nonparametric estimation of the response probabilities. The aim of this paper is to investigate, via simulation experiments, the small-sample properties of kernel regression estimation of the response probabilities when the auxiliary information consists in a mix of continuous and discrete variables. Furthermore the practical behavior of the method is evaluated on data of a web survey on accommodation facilities in the province of Florence.

1 Introduction

It is well known that unit nonresponse is a common problem in sample surveys. It has the potential to prevent valid inference especially when the group of respondents differs from that of non respondents. The standard way to attempt compensating for unit nonresponse is by some form of weighting adjustment. The essence of all weighting adjustment procedures is to increase the weights of the respondents so that they represent the nonrespondents. The respondents weights are increased by incorporating estimates of the probabilities that sampling units respond to the survey. The nonresponse is here viewed as an additional phase of sampling whose probability mechanism is unobserved and that follows the first phase of sampling, which is determined by the original sample design. A critical step of the weighting adjustment procedures is the estimation of the response probabilities. This step is usually conducted under explicit or implicit models relating response occurrences to auxiliary variables. The unspecified assumption here is that the response mechanism is missing at random (MAR). A common way to estimate the individual response probabilities (or weights) consists in partitioning the sampled units

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_16,

in "weighting classes" assumed homogeneous with respect to the mechanism of response and then in estimating the response probabilities as rates of respondent units within each class. This method, known as weighting within cell method, is based on the implicit model that assumes uniform response mechanism within subpopulations. Moreover sometimes it may be difficult to establish proper grouping, and incorrect grouping may result in bias or loss of efficiency (see Da Silva and Opsomer 2004). Another popular way to estimate the individual response probabilities is by fitting a parametric model such as logistic, probit or exponential model. An alternative solution can be the estimation of the response probabilities by nonparametric methods which require only that the response probabilities be related to the auxiliary variables by a smooth but unspecified function. The main motivation to use such methods is that they offer an appealing alternative to the choice of the link function, or when the parametric model is difficult to specify a priori. The use of kernel-type smoothing methods for estimating response probabilities was first proposed by Giommi (1984), then also considered by Giommi (1987) and by Niyonsenga (1994, 1997) and recently further investigated by Da Silva and Opsomer (2006, 2009) who theoretically derived some new important asymptotic properties of the non-parametrically weighting adjusted estimators and empirically evaluated their small-sample properties. They considered only the case in which a unique continuous covariate is available but their estimators can be extended with minor modifications to more auxiliary variables. However if some of the auxiliary variables are not continuous some assumptions required for the theoretical derivation of the asymptotic properties are not satisfied anymore. To evaluate the validity of the small sample properties in such situation, this paper investigates via simulation experiments the small-sample properties of kernel regression estimation of the response probabilities when the auxiliary information consists in a mix of continuous and discrete variables. The article is organized as follows. Basic definitions and assumptions are set out in Sect. 2. The performances of two kernel-type weighted mean estimators are discussed in Sect. 3. Section 4 concludes with final comments and ongoing questions.

2 Methodological Framework and Basic Assumption

Let U be a target population of N units labelled k $(k = 1, \ldots, N)$; let y be a study variable, of which we want to estimate the mean $\bar{Y} = \sum_{k \in U} y_k / N$ from a sample s of n units drawn from U according to a sampling design $p(s)$; and let π_k be the inclusion probability of unit k for all $k \in U$. When nonresponse occurs, the y_k's values are observed only on a subset of s, $r \subseteq s$ of $n_r (n_r \leq n)$ respondents, and it become necessary to model the response process in order to account for the loss in information in the estimation process. Unlike in the sampling selection, the survey sampler has no control over the response mechanism. Usually in order to model it, a response indicator R_k, assuming value one if the unit $k \in U$ responds and zero otherwise, is defined. The distribution of the vector $(R_k : k \in s)$ is called response

mechanism. In this paper we assume that the response mechanism is MAR and that a vector of auxiliary variables $\mathbf{x}$, correlated both to nonresponse and to y, is fully observed throughout the sample. Moreover we assume that, given the sample, the response indicators are independent random variables with:

$$\Pr(R_k = 1 | k \in s, y, \mathbf{x}) = \phi(\mathbf{x}_k) \equiv \phi_k, \quad \text{for all } k \in U \tag{1}$$

where the exact form of $\phi(.)$ is unspecified, but it is assumed to be a smooth function of $\mathbf{x}_k$, with $\phi(.) \in [0, 1]$. If all the response probabilities were conceptually known, the theory of two phase sampling lead us to the possible mean estimator:

$$\hat{\bar{y}} = \sum\nolimits_{k \in s} y_k \pi_k^{-1} \phi_k^{-1} R_k / \sum\nolimits_{k \in s} \pi_k^{-1} \phi_k^{-1} R_k \tag{2}$$

that is the Hàjek estimator adjusted to compensate for nonresponse (Little and Rubin 2002). In practice, however, this formula is unfeasible as the response probabilities are unknown. Then a possible estimator of the population mean is obtained replacing in (2) the response probabilities ϕ_k with their estimates $\hat{\phi}_k$, satisfying $0 < \hat{\phi}_k \leq 1$. The resulting weighting adjusted mean estimator is:

$$\hat{\bar{y}} = \sum\nolimits_{k \in s} y_k \pi_k^{-1} \hat{\phi}_k^{-1} R_k / \sum\nolimits_{k \in s} \pi_k^{-1} \hat{\phi}_k^{-1} R_k \tag{3}$$

In order to implement (3), it is necessary to estimate the response probabilities ϕ_k. For the case in which the auxiliary information consists in only one continuous variable Da Silva and Opsomer (2009) used kernel polynomial regression. Here we assume that a vector of q ($q \geq 2$) auxiliary variables of different type is fully observed throughout the sample and we extend the procedure of Da Silva and Opsomer (2009) to this situation. More in detail we consider two methods of estimating the response probabilities $\hat{\phi}_k$: the local constant and the local linear regression which result from the local fit of polynomials of degree zero and one respectively. Polynomials of greater degree are not considered because as noted in the literature (e.g. Fan and Gijbels 1996, p. 77) and also confirmed in the simulations of Da Silva and Opsomer (2009), higher degree polynomials reduce the bias but increase the variance. Both the local constant and local linear regression used are based on a "generalized product kernel function" $K(.)$, i.e. the product of q univariate kernel functions, one for each auxiliary variable. Each univariate kernel function may be a continuous data kernel function or a categorical ordered data kernel function or a categorical unordered data kernel function. For lack of space the detail of the estimation procedure are not given here and we refer to Li and Racine (2006) for a general description of local polynomial regression with mixed data. In this paper we extend the nonparametric techniques available in non survey literature to the inference for finite populations in the presence of nonresponse.

3 Simulation Study and Real Data Application

Some Monte Carlo experiments are performed in order to empirically evaluate the finite sample properties of the local constant and the local linear estimators of response probabilities. The simulations are carried out in R, and, for the nonparametric estimation procedures the R package "np" (Hayfield and Racine 2008) is used. Three different populations U_1, U_2, U_3, and two response mechanisms are generated. All the three populations are of 2,000 units and are partitioned by an indicator variable, x_2, in two subpopulations, each of 1,000 unit. In population U_1 the study variable y is related to a unique auxiliary variable $x_1 \sim Uniform(0,1)$ by the following linear model $y = 20 + 60x_1 + \varepsilon_1$, where $\varepsilon_1 \sim N(0,16)$. In population U_2 the parameter of the linear model relating the study variable to the variable x_1 is different in the two subpopulations, following the model $y = 20 + 60x_1x_2 + 45x_1(1-x_2) + \varepsilon_2$ where $\varepsilon_2 \sim N(0,9)$. In U_3, as in U_2, the study variable is related both to the continuous variable x_1 and to the categorical unordered variable x_2 but the relation with x_1 is quadratic; and the model is $y = (40 + 35x_1^2)x_2 + (50 + 30x_1^2)(1-x_2) + \varepsilon_3$ where $\varepsilon_3 \sim N(0,4)$. The response probabilities are generated, following the same procedure of Da Silva and Opsomer (2006), under the two response function:

$$\text{A}: \phi(\mathbf{x}) = (-\beta_0 x_1^2 + \beta_1 x_1 + \beta_2)x_2 + (-\beta_3 x_1^2 + \beta_4 x_1 + \beta_5)(1-x_2) \tag{4}$$

$$\text{B}: \phi(\mathbf{x}) = (e^{\beta_6+\beta_7 x_1}(1+e^{\beta_6+\beta_7 x_1})^{-1})x_2 + (e^{\beta_8+\beta_9 x_1}(1+e^{\beta_8+\beta_9 x_1})^{-1})(1-x_2) \tag{5}$$

where the coefficients $\beta_0, \ldots, \beta_9$, are chosen so that the response rate on the whole population is approximately equal to 0.7 and in the two subpopulations are approximately equal to 0.65 and 0.75.

For each of the six possible combinations of population models and response mechanisms, the simulation procedure consists of the following steps:

1. Select a simple random sample of 200 units.
2. Perform a Bernoulli trial for each unit $k \in s$ with probability ϕ_k(under model A or B) for "success" (response) and $(1-\phi_k)$ for "failure" (nonresponse).
3. Compute on the set of respondents the weighted mean estimators T_0, T_1, T_2, T_3, T_4, T_5 adjusted respectively with:
   - T_0 : true response probabilities
   - T_1 : response probabilities estimated through a logistic model
   - T_2 : response probabilities estimated through local linear regression
   - T_3 : response probabilities estimated through local constant regression
   - T_4 : response probabilities estimated through weighting within cell method
   - $T_5 : \hat{\phi}_k = 1, k \in s$ (naive estimator)
4. Repeat steps 1–3 3,000 times.

In all the four response probability estimation methods the auxiliary variables x_1 and x_2 are both used. In weighting within cell method the variables x_1 is categorized using the quartiles of its distribution and eight classes are defined crossing

the categorized x_1 with x_2. For both the local constant and the local linear regression procedure, the second order Epanechikov kernel function is used for x_1 and the Aitchison and Aitken kernel function (Aitchison and Aitken 1976) is used for x_2. In these last two methods two more choices in estimator settings concern the typology and the selection criterion of the bandwidth: the fixed type bandwidth and the least-squares cross-validation selection method are applied. The experimental results are reported in Tables 1 and 2 (with best performances in bold). Table 1 gives for each combination of population and response mechanism the percentage bias obtained with every adjustment procedures. Among the estimators affected by the generated nonresponse, the worst bias performances are obviously those for the unadjusted "naive" estimator. Biases are successfully reduced with all the other estimators. Apart from the estimator adjusted by the true response probabilities, that is conditionally unbiased for the full sample estimates, the best bias reduction is obtained by the weighting within cell adjustment method for the response mechanism A and by the local linear regression adjustment method for the response mechanism B. The local constant regression adjustment method produces for all the populations and both the response mechanisms the worst bias reduction with respect to all the other response probability estimation methods. The logistic adjustment method is better than the weighting within cell adjustment method for the response mechanism B.

Table 2 shows the empirical mean square error (MSE) of the applied adjustment methods. The best performances in terms of smallest MSE are obtained using the local linear regression method in four of the six cases and with the logistic adjustment method in the other two with a slightly difference between the two methods. Both the two local regression adjustment methods and the logistic

Table 1 Percentage biases of weighting adjusted mean estimator

Response functions	Populations	T_0	T_1	T_2	T_3	T_4	T_5
A	U_1	0.006	0.267	0.134	0.703	**0.033**	2.600
A	U_2	0.020	−0.349	0.195	0.363	**−0.153**	1.327
A	U_3	0.022	0.194	−0.179	0.395	**0.065**	1.842
B	U_1	−0.047	0.227	**−0.173**	−1.286	−0.737	−7.322
B	U_2	−0.023	0.069	**0.043**	−0.514	−0.435	−3.137
B	U_3	−0.068	0.648	**−0.307**	−1.320	−0.662	7.249

Table 2 Empirical mean square error of weighting adjusted mean estimator

Response functions	Populations	T_0	T_1	T_2	T_3	T_4	T_5
A	U_1	2.448	**1.486**	1.528	1.690	3.950	3.632
A	U_2	0.862	0.579	**0.567**	0.604	1.327	1.280
A	U_3	2.025	**1.227**	1.248	1.327	2.529	2.312
B	U_1	2.362	1.551	**1.532**	2.026	4.076	15.366
B	U_2	0.948	0.605	**0.572**	0.670	1.396	3.724
B	U_3	1.933	1.389	**1.273**	1.694	2.631	12.833

adjustment method produce estimators more efficient than those adjusted by the true response probabilities. The weighting within cell adjustment method produces the greatest MSE with respect to all the other response estimation methods and it produces a MSE even slightly greater than that of the naive estimator under response mechanism A.

The practical behavior of the two non parametrically adjusted estimators is also evaluated through an application on real data. The data derive from a web survey carried out monthly by the Province of Florence in order to obtain timely data on tourist movement in the accommodation facilities located in the province area. The survey collects information on the number of tourist that daily arrive in each accommodation facilities, the number of permanence days for each tourist and some characteristic of the accommodation facilities and of the tourists. It is based on an auto-selected sample of the accommodation facilities that agreed to participate to an automatic system of registration and transmission of data via web. The participation rate is very low: In year 2006, to which our data refer, the average response rate on the 12 months is about 21%. The web survey data are used to estimate the total number of nights spent per month. To estimate the response probabilities three variables are used: the number of beds, the number of stars and a "quality index" defined as the ratio between the number of bathrooms and the number of rooms. The estimates are then compared with the total evaluated using the data of a census survey carried out by the Italian Statistical Institute (ISTAT) which are assumed as true values. The ISTAT survey collects the same information of the web survey but it is a national monthly census survey based on a complex system and the data are available only after some months. The results are very interesting: the monthly mean relative bias is over 30% using the naive estimator but it decreases to 3.9% and 4.3% considering the weighting adjusted mean estimators with response probabilities estimated respectively through the local linear and the local constant regression. The monthly mean relative bias of the logistic adjustment method is about 10%. The weighting within cell adjusting method is not considered due to the difficulty in identifying the weighting classes.

4 Concluding Remarks and Ongoing Questions

The results of both the simulation experiments and the application to real data show that the two nonparametric weighting adjusted estimators examined could be competitive, in terms of bias and MSE, with respect to the weighting adjusted estimators with response probabilities estimated both through a logistic parametric model and by the weighting within cell method.

There are still a number of open questions that need to be addressed. First, in the simulation study we consider only six scenarios in terms of populations and response functions and more investigations are needed to validate the promising results. Second, in the nonparametric approach the exact form of $\phi(.)$ is unspecified but some estimator settings (like the form of the kernel functions, the type and

the selection criterion of the bandwidth) are still required: we followed the common settings applied in literature, however some further investigation on the effects of different settings could be interesting. In addition, we also plan to estimate the variance of the mean estimators and to analyze the conditions for which a weighting adjusted mean estimator based on estimated response probabilities may result more efficient than the corresponding estimator based on true response probabilities. Some authors, including Rosenbaum (1987), Robins et al. (1994), and Little and Vartivarian (2005) just noted that estimators using the estimated response probabilities can be more efficient than the corresponding estimators using the true response probabilities. Beaumont (2005) gave a clear justifications for the variance reduction obtained using estimated response probabilities from a logistic regression model in the imputation context. Kim and Kim (2007) extended the Beaumont's results to all the cases in which the parameters of the response probabilities are estimated by the maximum likelihood method. Based on such literature and on the analysis of the correlations existing in our simulated data between the study variable and the covariates and between the study variable and the true response probability, we think that the adjustment using the estimated response probability improves efficiency if the auxiliary variables used in the response estimation procedure include additional information on the study variable and if the procedure used to estimate the response probabilities is able to incorporate this additional information.

Acknowledgements The work was supported by the Italian Ministry of University and Research, MIUR-PRIN2007: Efficient use of auxiliary information at the design and at the estimation stage of complex surveys: methodological aspects and applications for producing official statistics.

References

Aitchison, J., & Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrica, 68*, 301–309.

Beaumont, J. F. (2005). Calibrated imputations in survey under a model-assisted approach. *Journal of the Royal Statistical Society Series B, 67*, 445–458.

Da Silva, D. N., & Opsomer, J. D. (2004). Properties of the weighting cell estimator under a nonparametric response mechanism. *Survey Methodology, 30*, 45–55.

Da Silva, D. N., & Opsomer, J. D. (2006). A kernel smoothing method of adjusting for unit nonresponse in sample survey. *The Canadian Journal of Statistics, 34*, 563–579.

Da Silva, D. N., & Opsomer, J. D. (2009). Nonparametric propensity weighting for survey nonresponse through local polynomial regression. *Survey Methodology, 35*, 165–176.

Fan, J., & Gijbels, I. (1996). *Local polynomial modeling and its applications*. London: Chapman & Hall.

Giommi, A. (1984). A simple method for estimating individual response probabilities in sampling from finite populations. *Metron, 42*, 185–200.

Giommi, A. (1987). Nonparametric methods for estimating individual response probabilities. *Survey Methodology, 13*, 127–134.

Hayfield, T., & Racine, J. S. (2008). Nonparametric econometrics: the np package. *Journal of Statistical Software, 27*(5). URL http://www.jstatsoft.org/v27/i05/.

Kim, J. K., & Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probabilities. *The Canadian Journal of Statistics, 35*, 501–514.

Li, Q., & Racine, J. S. (2006). *Nonparametric econometrics: Theory and practice*. Princeton, NJ: Princeton University Press.

Little, R. J. A., & Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey mean. *Survey Methodology, 31*, 161–168.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: Wiley.

Niyonsenga, T. (1994). Nonparametric estimation of response probabilities in sampling theory. *Survey Methodology, 20*, 177–184.

Niyonsenga, T. (1997). Response probability estimation. *Journal of Statistical Planning and Inference, 59*, 111–126.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association, 89*, 846–866.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association, 82*, 387–394.

Part IV
Data Analysis in Industry and Services

Measurement Errors and Uncertainty: A Statistical Perspective

Laura Deldossi and Diego Zappa

Abstract Evaluation of measurement systems is necessary in many industrial contexts. The literature on this topic is mainly focused on how to measure uncertainties for systems that yield continuous output. Few references are available for categorical data and they are briefly recalled in this paper. Finally a new proposal to measure uncertainty when the output is bounded ordinal is introduced.

1 Introduction

In measurement system analysis (MSA) the assessment of the measurement errors is a very relevant topic. Especially in the industrial context it allows to understand whether the differences among the results of an experiment or the variability of a process are due to either an out of control process or to a non capable measurement system. In the last 15 years the International Organization of Standardization (ISO) has produced several documents on this topic. The most important international standards are the ISO 5725:1994 (Parts 1–6) (International Organization for Standardization 1994) and the "Guide to the expression of Uncertainty in Measurement" (GUM) published in 1995 (International Organization for Standardization 1995), which propose quite different approaches to the validation of the measurement procedure, though they were published almost in the same year (see Deldossi and Zappa 2009, for a comparison). The main distinction is related to the topic of interest: accuracy in ISO 5725, uncertainty in GUM. While the term accuracy is related both to trueness and precision, in GUM the relevance is focused only to the uncertainty (precision) aspect.

At present the statistical literature on MSA is mainly aimed to assess measurement system precision. For this reason our contribution will concern only this topic, omitting other aspects of measurement system, such as calibration or assessing linearity. In Sect. 2 we classify the approaches available in statistical literature to evaluate "precision" for measurement systems yielding both continuous and categorical output. In Sect. 3 a new proposal to measure uncertainty for ordinal data

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_17,

based on CUB models (D'Elia and Piccolo 2005) is introduced, followed by an example, some comments and additional details.

2 Methods to Measure Precision: A Classification

The precision of a measurement system concerns its consistency across multiple measurement per object. Generally, using experimental designs, data are collected on a sample set of n objects in the following manner: each object i (for $i = 1, \ldots n$) is evaluated once or t times by m appraisers; the measurement value that is assigned to object i by appraiser j (for $j = 1, \ldots, m$) in the $h-$th measurement (for $h = 1, \ldots, t$) will be labelled Y_{ijh}. In case of measurements on a categorical scale, Y_{ijh} is assigned by choosing a category within a finite, bounded, ordered or unordered set $\{1, 2, \ldots, a\}$. Otherwise, when the measure is defined on a metric scale, Y_{ijh} will assume values on a continuous interval. Correspondingly, let $X_1, X_2, \ldots, X_n$ be the true values of the n objects. These latent values can be defined on a scale different from the one used for the related observed values. For example, the true value underlying a categorical scale does not need to be categorical. However, if the true value is continuous, this may cause an ordering in the categories and then one would expect the observed value will be expressed at least on an ordinal scale. Short details on methods used to evaluate precisions are reported in the following.

(a) Suppose Y_{ijh} is on a metric scale. The standard method to assess the precision is the gauge repeatability and reproducibility (GR&R) experiment (Burdick et al. 2003), based on the concept of R&R variances. These are typically estimated by a random effect model, like:

$$Y_{ijh} = \mu + P_i + O_j + (PO)_{ij} + E_{ijh} \tag{1}$$

where μ represents the overall mean, P_i, O_j, $(PO)_{ij}$ and E_{ijh} are supposed to be jointly independent normal random variables (r.v.) with mean zero and variance σ^2_P, σ^2_O, σ^2_{PO} and σ^2_E, respectively. P and O generally stand for "Part" and "Operator". The gauge variance is equal to $\sigma^2_O + \sigma^2_{PO} + \sigma^2_E$ as the measurement system generally refers both to the instrument used and its interaction with the operators who have used it. ISO 5725 is mainly based on this method. Suitable estimators of variance components are obtained by standard ANOVA table (see Burdick et al. 2003).

(b) Suppose Y_{ijh} is on a categorical scale. The measurement system may be defined on either (b1) nominal or (b2) ordinal scale. Typical contexts are related to quality inspectors, helpdesk employees or physicians, who must classify or judge the quality of some objects. It is obvious that in such a context there are no instruments and only subjective ability and/or training of the appraisers may contribute to the capability of the measurement system.

(b1) For nominal categorical scale, reproducibility makes reference to the probability of agreement, P_A. Assuming henceforth that each object is measured once by each appraisers j, P_A represents the probability that two arbitrary measurements of an arbitrary object, i, are identical, that is,

$$P_A = P(Y_{ij1} = Y_{ij2}) = \sum_{l=1}^{a} \sum_{k=1}^{a} p(l) q^2(k|l) \tag{2}$$

where a is the number of categories that appraisers j_1 and j_2 can assign to object i; $p(l) := P(X_i = l)$, for $l = 1, \ldots, a$, is the discrete probability distribution of the latent variable X_i; $q(k|l) := P(Y_{ij} = k|X_i = l)$, for $k, l = 1, \ldots, a$, is the distribution of the measurement error. Notice that two appraisers would agree even if they assign values to objects at random. To deal with this problem, Cohen (1960) and several other authors introduced the so-called K-type indexes as a rescaled version of P_A. In particular the K-index introduced by De Màst and Van Wieringen (2007) is:

$$K = \frac{P_A - P_{A|chance}}{1 - P_{A|chance}} \tag{3}$$

where $P_{A|chance}$ is the probability of agreement for a completely uninformative measurement system that randomly assigns measurement values to objects. If we adopt the uniform distribution when the measurement system works at random, (3) becomes:

$$K^{UNIF} = \frac{P_A - 1/a}{1 - 1/a}. \tag{4}$$

It follows that the relevant range of K^{UNIF} is [0,1], while P_A ranges within $[1/a, 1]$, then depending on the number of categories. K^{UNIF} is a precision index that allows comparisons among different nominal measurement systems. An unbiased estimator of (2) and (4) is reported in De Màst and Van Wieringen (2007).

(b2) If Y is a bounded ordinal variable, we have to distinguish between the cases in which the latent variable X is continuous or categorical. In the first one a discretization is necessary, since the ordinal variable Y cannot be directly modeled as $Y = X + E$, i.e. as the sum of the two continuous variables X and E. De Màst and Van Wieringen (2004) propose to link the observed and the latent variable by:

$$Y = \text{LRD}\,(X + E) = \left\lceil \frac{a \cdot \exp(X + E)}{1 + \exp(X + E)} \right\rceil \tag{5}$$

where $\lceil \cdot \rceil$ is the upper integer operator. Inverting (5) and taking into account the necessity to avoid indeterminate results when $y = 0$ or $y = a$, we have

$$X = \mathrm{LRD}^{-1}(Y) = \ln\left(\frac{Y - 1/2}{a - Y + 1/2}\right). \tag{6}$$

Once Y is transformed by (6), they propose to evaluate the precision of the measurement system using GR&R indices on X as it was a continuous variable. In particular they propose to use the intraclass correlation coefficient $\rho_X = \sigma_X^2/\sigma_Y^2$. The higher ρ_X, the higher the precision of the measurements will be.

At present, the literature lacks of proposals for evaluating the measurement system capability when the latent variable X is categorical and Y ordinal. Only non-parametric indexes, based on ranks such as Kendall-τ and Spearman-ρ are available (see De Màst and Van Wieringen 2004), but no suggestions on how to measure uncertainty is given.

3 A New Proposal to MSA for Ordinal Data

In case of Y is bounded ordinal, we propose an approach to ascertain the measurement precision exploring the potentials of the CUB model, presented by D'Elia and Piccolo (2005). Authors have introduced and used this model to explain the behavior of respondents facing with judgments (rating). It is defined as a mixture of a shifted Binomial and a discrete Uniform random variable as it follows

$$P(Y = k) = \pi \binom{a-1}{k-1} (1-\xi)^{k-1} \xi^{a-k} + (1-\pi)\frac{1}{a} \tag{7}$$

with $\pi \in (0, 1], \xi \in [0, 1], k \in \{1, 2, \ldots, a\}$. The first instance of a mixture model for ordinal data and the inferential issues are derived by Piccolo (2003, 2006) respectively. For the matter of parameter identifiability in (7) it must be $a > 3$. In our context the Uniform component may express the degree of uncertainty in judging an object on the categorical scale $\{1, 2, \ldots, a\}$, while the shifted Binomial random variable may represent the behavior of the rater with respect to the liking/disliking feeling towards the object under evaluation: the parameter ξ is then a proxy of the rating measure, while π is inversely related to the uncertainties in the rating process. The statistical properties of (7), the E-M algorithm needed to find the estimates of π and ξ, the extensions on how to exploit covariates linked to the parameter space, as well as numerical issues, are described in Iannario and Piccolo (2009). In addition we have prepared an independent tool in Excel (freely available) that, given a sample set, computes the π and ξ estimates, without using the E-M algorithm, but searching for the maximum of the likelihood, having preliminarily found the subdomain where the desired solution is placed. To give an idea, let $a = 4$ and $m = 6$ and suppose to have observed the sample set $\{2, 2, 2, 1, 2, 4\}$, gathering appraisers' evaluations. Figure 1 reports the related likelihood surface. Arrows show where the maximum is expected to be placed.

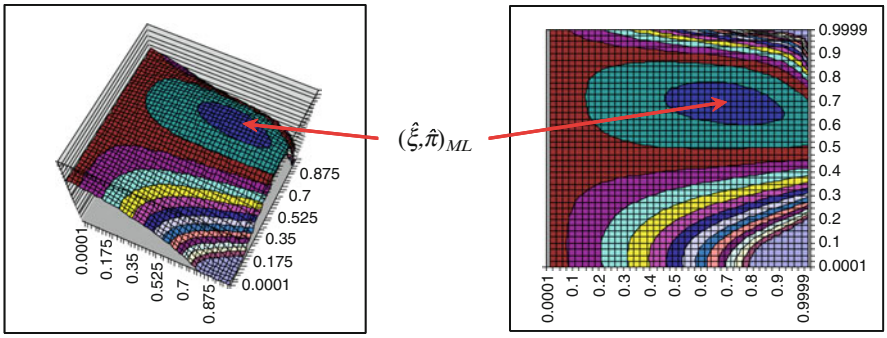


Fig. 1 The likelihood surface for (7): an example for $a = 4$ and $m = 6$

The solutions found with our procedure slightly differs from the ones we get with R (it is a matter of the degree of convergence we fix in our program) and of course the computer time is longer. The reason why we have prepared this application is to share results also with those (potentially not few) industrial contexts where R is at present still unknown and some training would be needed.

In MSA studies the efficacy of the CUB model stands in the possibility to separate the evaluation of a part either good, bad or of intermediate quality from the uncertainty that appraisers have in evaluating that part.

The main interest is on the level of π: the smaller is π, the stronger is the measurement uncertainty. For a matter of interpretation, we may say that for a given object

$\pi \cong 0$ if the appraisers choose at random a value in $\{1, 2, \ldots, a\}$
$\pi = 1$ if all the appraisers give the same rating to the objects.

However some problems exist in the application of CUB in MSA. First of all the number of the categories must be $a > 3$: this constrain does not allow the application of CUB to binary data. Furthermore, inferential properties described in D'Elia and Piccolo (2005) are based on asymptotic results. In MSA the number of appraisers (i.e. the sample size used in CUB) is small (the number, m, of appraisers is often not greater than 10). As a consequence, we cannot apply any asymptotic results and we had to investigate on the implications.

An advantage of dealing with a small sample set should be that the parameter space is a finite set. Given m and a, the number of possible combinations with replication (i.e. the possible sequence of responses to a given object by m appraisers) is $\binom{m+a-1}{m}$. For example, if $a = 4$ and $m = 6$ the cardinality of the parameter space is 84. Then, for inferential purposes we may find (exactly) all the feasible estimates. However, if either a or m increase, the cardinality may be considered a countable infinite set and some computational problems may arise. To make our procedure the most general as possible, we have implemented in our spreadsheet a bootstrap procedure to compare the distribution of the estimator based on the

available data set with the one we obtain by assuming the appraisers are evaluating objects at random.

When dealing with small m, a problem may arise regarding the interpretation of the π parameter because it turns out to be bounded from below. Table 1 reports the lowest estimate of π corresponding to the maximum uncertainty for some combinations of a and m.

From Table 1 it is clear that we cannot approach 0 when $m \neq q{\cdot}a$ for $q = 1, 2, 3 \ldots$, as the maximum heterogeneity, which should assure $\pi \to 0$, cannot be achieved. The bounds have been calculated supposing to distribute the judgements made by m appraisers in order to reproduce a configuration of responses closed to maximum uncertainty. The matter is that the solution may not be unique. For example, for $a = 4$ and $m = 5$ we may suppose that four appraisers choose respectively the value $y = 1$, $y = 2$, $y = 3$, $y = 4$ and the last one an arbitrary value between 1 and 4. Then we have four different configurations of "quasi" maximum uncertainty and for each of them we have computed $\hat{\pi}$ reporting in Table 1 the minimum estimate.

To exemplify most of the details introduced above and for a matter of comparison with other studies, consider the data set available in De Màst and Van Wieringen (2004), related to a real database concerning a printer assembly line. Some records are reported in Table 2. After a printer has been assembled its quality is tested by printing a grey area. The latter is visually inspected on uniformity by six operators who give a quality evaluation using the ordinal scale 1:good, 2:acceptable, 3:questionable, 4:rejected.

In this example $a = 4, m = 6$. We have estimated the couple $(\hat{\pi}_i, \hat{\xi}_i)$ for each object $i = 1, \ldots, 26$. The scatterplot of the estimates is in Fig. 2.

It may be observed that for some printers the appraisers had no uncertainty in the evaluation ($\hat{\pi} = 1$) while for some other printers, $\hat{\pi} = 0.1112$, which is, according to Table 1, the value corresponding to the maximum uncertainty. In order to assess whether the mean of the 26 $\hat{\pi}$ estimates, $\bar{\pi} = 0.6027$, is small (or large), Fig. 3 plots the cumulative bootstrap distribution, ${}_B\bar{\pi}_{MU}$, assuming that 6 appraisers have

Table 1 Lowest bound of $\hat{\pi}$ for some combination of a and m

a \ m	4	5	6	7	8	9	10	11	12
4	0	0.2000	0.1112	0.1125	0	0.1112	0.0667	0.0723	0
5	0.1099	0	0.0706	0.1071	0.0625	0.0490	0	0.0413	0.0625
6	0.1000	0.0625	0	0.0708	0.1000	0.0667	0.0400	0.0284	0

Table 2 Some records from the printer assembly data (De Màst and Van Wieringen 2004)

		Operators					
		1	2	3	4	5	6
Objects	1	4	4	4	4	2	4
	2	1	4	2	3	3	4
	…	…	…	…	…	…	…
	26	1	1	2	2	4	4

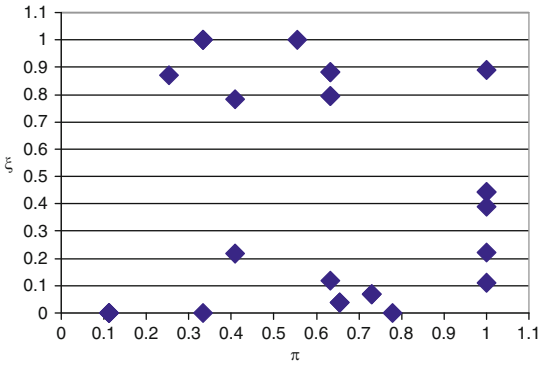


Fig. 2 Scatterplot of $(\hat{\pi}_i, \hat{\xi}_i)$ for the printer assembly data

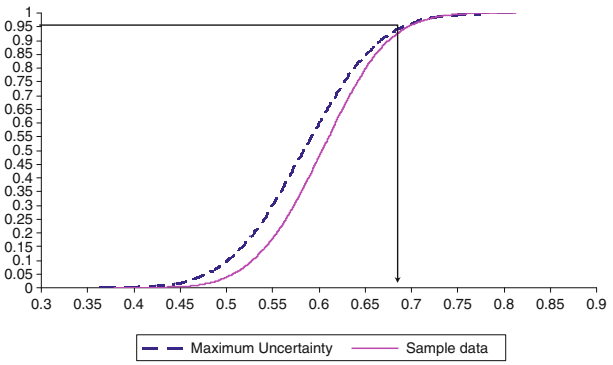


Fig. 3 Cumulative bootstrap distribution of ${}_B\bar{\pi}$ (*solid line*) and ${}_B\bar{\pi}_{MU}$ (*dashed line*)

evaluated 26 objects at random, i.e. replicating a condition of maximum uncertainty. For comparison the cumulative bootstrap distribution, ${}_B\bar{\pi}$, starting from the sample set is also reported.

The arrow in Fig. 3 shows where the 95th percentile of ${}_B\bar{\pi}_{MU}$ is placed. It is clear that it is on the right of $\bar{\pi}$, supporting the hypothesis that measurement system is not capable of evaluating the printer assembly data. If we had used the indexes described in Sect. 2 (with the exception of ρ_X as it assumes the existence of a continuous latent variable - an assumption not appropriate for this example), we should have found

$$\hat{\mathrm{P}}_A = 0.302564, \hat{K}^{UNIF} = 0.07008.$$

Both the estimates give an evidence of a non capable measurement system. In particular $\hat{K}^{UNIF}$, which considers the possibility of answering at random, is very closed to 0, underlining that the appraisers do not agree with the evaluation of the printers.

Table 3 Parameter estimates of uncertainty by operator

Operator j	$\hat{\pi}_j$	$\hat{\xi}_j$	p-Value
1	0.4337	0.9122	0.245
2	0.1795	1E-08	0.595
3	0.7282	0.7073	0.054
4	0.3195	0.2862	0.381
5	0.4086	0.0503	0.261
6	0.6923	1E-08	0.069

Additional information can be gathered using the CUB model. On one hand we may measure the precision by evaluating its "distance" from the worst scenario, assuming all the appraisers evaluate objects at random (greater is the distance, better is the measurement precision); on the other hand we may also analyze the reason of a bad performance of the measurement system. For example, we may assess whether uncertainty is uniformly shared by all the appraiser, reflecting the (equally perceived) difficulty in judging the objects. In fact we may exploit CUB models assuming that we are interested in measuring the uncertainty of the appraisers in evaluating a sample of $m = 26$ objects. Table 3 reports the $(\hat{\pi}_j, \hat{\xi}_j)$ estimates, for $j = 1, 2 \ldots 6$. The last column is the p-value of $\hat{\pi}_j$ refereed to $\Pr(\hat{\pi}_{MU} > \hat{\pi}_j)$, where the distribution of $\hat{\pi}_{MU}$ has been found by simulating the event of giving at random a category $a = 1, \ldots, 4$ to each object i (for $i = 1, \ldots, 26$).

It may be observed that operators #3 and #6 had the least uncertainty in evaluating the objects, but they had quite different perception of the overall quality ($\hat{\xi}_3 = 0.7$, $\hat{\xi}_6 = 0$). Analogously, comments (maybe different) can be done for the other operators. Then a reason of the overall uncertainty may be addressed both to a non-homogenous sample set and to a non equally trained set of operators.

These conclusions could be refined applying the extended version of CUB models reported in Iannario and Piccolo (2009) taking into account covariates on the operators (e.g. gender, age, experience,...).

References

Burdick, R. K., Borror, C. M., & Montgomery, D. C. (2003). A review of methods for measurement systems capability analysis. *Technometrics*, 43, 342–354.

Cohen, J. (1960). Coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.

De Màst, J., & Van Wieringen, W. N. (2004). Measurement system analysis for bounded ordinal data. *Quality and Reliability Engineering International*, 20, 383–395.

De Màst, J., & Van Wieringen, W. N. (2007). Measurement system analysis for categorical measurements: agreement and kappa-type indices. *The Journal of Quality Technology*, 39, 191–202.

Deldossi, L., & Zappa, D. (2009). ISO 5725 and GUM: comparison and comments. *Accreditation and Quality Assurance*, 3, 159–166.

D'Elia, A., & Piccolo, D. (2005). A mixture model for preference data analysis. *Computational Statistics and Data Analysis*, 49, 917–934.

Iannario, M., & Piccolo, D. (2009). *A program in R for CUB models inference, Version 2.0*. Available at http://www.dipstat.unina.it

International Organization for Standardization (ISO) (1994). ISO 5725. Geneva, Switzerland: ISO.

International Organization for Standardization (ISO) (1995). *Guide to the expression of uncertainty in measurement*. Geneva, Switzerland: ISO.

Piccolo, D. (2003). On the moments of a mixture of uniform and shifted binomial random variables. *Quaderni di Statistica*, 5, 85–104.

Piccolo, D. (2006). Observed information matrix for MUB models. *Quaderni di Statistica*, 8, 33–78.

Measurement Uncertainty in Quantitative Chimerism Monitoring after Stem Cell Transplantation

Ron S. Kenett, Deborah Koltai, and Don Kristt

Abstract Following allogeneic stem cell transplantation, graft status is often inferred from values for DNA chimerism in blood or bone marrow. Timely assessment of graft status is critical to determine proper management of post cell transplantation. A common methodology for chimerism testing is based on STR-PCR, i.e. PCR amplification of **S**hort **T**andem DNA **R**epeats. This is a complex technology platform for indirect DNA measurement. It is however associated with inherent variability originating from preparation, amplification of the DNA, and uncalibrated product detection. Nonetheless, these semi-quantitative measurements of DNA quantity are used to determine graft status from estimated percent chimerism [*%Chim*]. Multiplex PCR partially overcomes this limitation by using a set of simultaneously amplified STR markers, that enables computing a [*mean%Chim*] value for the sample. Quantitative assessment of measurement variability and sources of error in [*mean%Chim*] is particularly important for longitudinal monitoring of graft status. In such cases, it is necessary to correctly interpret differential changes *of* [*mean%Chim*] as reflective of the biological status of the graft, and not mere error of the assay. This paper presents a systematic approach to assessing different sources of STR measurement uncertainty in the tracking of chimerism. Severe procedural and cost constraints are making this assessment non trivial. We present our results in the context of Practical Statistical Efficiency (PSE), the practical impact of Statistical work, and InfoQ, the Information Quality encapsulated in ChimerTrack®, a software application tracking chimerism.

1 Introduction

The graft status following hematopoietic allogeneic stem cell transplantation is often inferred from values for DNA chimerism in blood or bone marrow, i.e. the ratio of donor to recipient DNA. Timely assessment of graft status is critical to determine proper management following transportation. DNA fingerprinting or DNA typing as it is known now was first described in 1985 by an English geneticist named Alec Jeffreys. Jeffreys found that certain regions of DNA contained DNA sequences that are repeated over and over again next to each other. He also discovered that the

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_18,

number of repeated sections present in a sample could differ from individual to individual and exhibit high polymorphism. By developing a technique to examine the length variation of these DNA repeated sequences, and then amplifying the quantity of these loci with PCR, Jeffreys created the ability to perform human identity tests. These repeated DNA regions became know as VNTR's, Variable Number of Tandem Repeats. An effective solution for DNA typing, including a high power of discrimination and rapid analysis, has been achieved with Short Tandem Repeat (STR), which has become the most quantifiable common platform for clinical chimerism monitoring. Multiple STRs from a single individual can be examined in the same DNA test, or 'multiplexed'. Multiplex STR examination is valuable because the platform is frought with intrinsic errors of DNA measurements. Using multiple STR loci from the same sample partially overcomes this limitation since it enables computing a [mean%Chim] value for the sample. In addition, the detection of multiplex STRs can be automated. However, STR analysis is a complex technology platform for indirect DNA measurement, associated with inherent variability originating from preparation, amplification of the DNA, and uncalibrated product detection. Nonetheless, these semi-quantitative measurements of DNA quantity are used to determine graft status from estimated percent chimerism [%Chim]. Quantitative assessment of measurement variability and sources of error in [mean%Chim] is particularly important for longitudinal monitoring of graft status. In such cases, it is necessary to correctly interpret differential changes of [mean%Chim] as reflective of the biological status of the graft, and not mere error of the assay.

Measurement systems, such as ChimerTrack®, are designed to support decision makers (Kristt et al. 2004, 2005, 2007; Kristt and Klein 2004). Proper assessment of a patient's graft status is of life critical importance. Physicians are faced with false positive and false negative outcomes which can have serious consequences for the patient being monitored. In order to evaluate decision support systems such as Chimertrack® we will refer to the quality of the information generated by the system. In this assessment, two fundamental consets will be used: Practical Statistical Efficiency (PSE) and Information Quality (InfoQ). PSE is evaluating the impact of a statistical procedure in a comprehensive way (Kenett 2007). InfoQ has been developed as a tool for assessing information gained from data and analysis (Kenett and Shmueli 2009). In this paper we assess the measurement uncertainty of the ChimerTrack® software and evaluate it's PSE and the level of InfoQ generated by it. The next section describes the chimerism measuring process, Sect. 3 discusses its measurement uncertainty and Sect. 4 presents the PSE and the InfoQ level of ChimerTrack®. We conclude with a summary and directions for future work.

2 The Chimerism Measurement Process

Measuring chimerism involves seven sequential steps described below. We focus on the potential causes of measurement uncertainty and the specific data sources that will be used in assessing the measurement uncertainty.

2.1 Taking Blood from the Patient

This is a regular blood taking process. Some of this blood is used for STR analysis, and some for traditional tissue typing evaluations

2.2 DNA Extraction

The blood sample consists of cellular and fluid components. The cells are isolated and their DNA must be separated from other cell constituents before the STR-DNA can be analyzed. DNA extraction methods have been developed for this purpose. This stage does not provide a numerical result, but rather it prepares the DNA for the following stage.

2.3 Purity and DNA Concentration Calculation

In order to proceed with the multiplex Polymerase Chain Reaction (PCR) process, the quantity and quality of extracted DNA has to be measured. At this stage the following data is calculated:

- The concentration of the DNA in the mixture received.
- The density of the material at a wave length of 260 and 280 nm.
- The relation between the density at a wavelength of 260 nm and density at a wavelength of 280 nm.

2.4 STR/PCR

Each STR marker is actually a set (or system) of many alleles all sharing the basic base structure of the repeat, but differing in the number of tandem repeats of this sequence. In the SGM Plus kit used in this work, there are 10 markers each with 8–23 different-sized alleles (Applied Biosystem 2006). An individual will normally have only one or two STR alleles in a marker system, depending on whether he is homozygous or heterozygous, respectively, at that marker. Similarly, a chimeric marker locus will have one to four allelic peaks (bands). In addition, for genetic and technical reasons, not all of these markers are useable for analysis, as explained below. The markers that are useable, or "informative", will be referred to as the marker profile for a patient, and will function as a personalized set of chimerism markers for all samples from a specific donor-recipient pair. It usually consists of 3–7 of the 10 marker sets in the SGM Plus kit (Kristt et al. 2005). PCR allows producing large amount of a DNA target nucleotide sequence by a process of cyclic

amplifications. The relatively large amount of resulting DNA enables analyzing the STR target sequences quantitatively.

2.5 *Electrophoresis*

DNA electrophoresis is an analytical technique used to separate the STR-DNA fragments by size. An electric field forces the fragments to migrate through a gel. DNA molecules normally migrate from negative to positive potential due to the net negative charge of the phosphate backbone of the DNA chain. Smaller DNA fragment migrate faster and further over a given period of time than do larger fragments.

2.6 *Genescan^{TM} Software*

PCR STR-DNA products are separated using a process of capillary electrophoresis, which provides two types of data: time to measurement and the intensity of the fluoresce tag on each STR-DNA molecule. These raw data are then transformed into more useful numerical data of peak area and molecular size of the STR marker in basepairs by using the ABI Genescan software The data that we receive at this stage consists of: (1) *Minutes* = the time of a molecule reaching one end from the other (the raw data), (2) *Size* = the size of the molecule calculated by using the minutes and (3) *Peak area* = the area that is under the allele peak which is calculated by using the minutes.

2.7 *ChimerTrack*®

This is a software application built on an Excel spreadsheet. By transferring data from Genescan to Excel, it is possible to automatically calculate the percent chimerism it's standard deviation, and measurement error (Kristt et al. 2004, 2005, 2007; Kristt and Klein 2004) In addition, the software displays the longitudinal course of a patient's chimeric status by creating a graph that tracks all of the patient's test results that were performed up to and including the present sample (Kristt et al. 2004). A sample ChimerTrack® report is presented in Fig. 1.

The chimerism of each marker is computed as $\boldsymbol{Chim} = \frac{(d_1+d_2)}{(d_1+d_2+r_1+r_2)}$, where d_1, d_2 are the peak area of the first and the second allele of the donor respectively, and r_1, r_2 are the peak area of the first and the second allele of the recipient, respectively. The ***[mean%Chim]*** is the mean of the chimerism of all the markers that have been used.

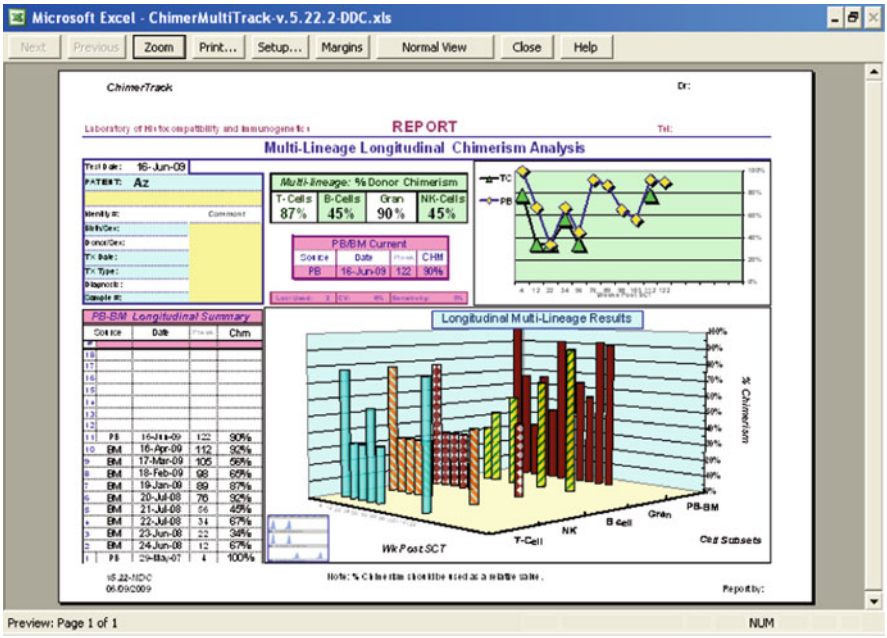


Fig. 1 ChimerTrack® output

$$\textit{[mean\%Chim]} = \frac{100 * \sum_{i=1}^{n} \left(\frac{d_{i1}+d_{i2}}{d_{i1}+d_{i2}+r_{i1}+r_{i2}} \right)}{n}$$ where n is the number of markers that have been used.

3 Measurement Uncertainty of Chimerism

In assessing the measurement uncertainty of chimerism, we combine an analysis of observational data with proactive experiments, specifically designed for assessing the measurement uncertainty. This mixed approach strategy was implemented in order to meet cost and practical constraints. The data used in the measurement of uncertainty includes the concentration of DNA the purity and the chimerism in 66 blood samples. This data was collected at the Rabin Medical Center in Petach Tikvah, Israel. In addition, an experiment was performed in cooperation with the Laboratory of Histocompatibility and Immunogenetics to assess the measurement error of the concentration measurement. In this experiment the purity and concentration of 20 DNA mixtures was calculated. The calculation for each mixture was repeated five times and covered a large range of results. A third database used in the measurement uncertainty assessment, consists of results of an experiment done on 38 pairs of DNA used for paternity testing mixtures. Each sample was amplified during the PCR phase and then inserted into the electrophoresis machine. The

data generated by electrophoresis was analyzed by the Genescan software. This data includes all the peak areas and sizes of each allele of each marker. Only heterozygote alleles with data from the two replications was included in our analysis. The replications were done on different PCR machines. In addition the relationship between concentration of DNA and chimerism was checked using data on the concentration and chimerism of 66 blood samples. Overall, the strategy we implemented combined careful mapping of the measurement process, as outlined in Sect. 2, with a thorough search for existing data and planning of feasible proactive experiments. This combined strategy is required for complex cases, such as chimerism, where textbook solutions cannot be implemented "as is".

In assessing the chimerism measurement uncertainty we combine several methods of analysis. First we analyze the relationship between concentration of DNA and chimerism. We then proceed with an analysis of the additional data sources. Finally we combine the results for an overall assessment of measurement uncertainty of *[mean%Chim*] In the next section we evaluate the ChimerTrack® decision support system from a more general perspective including PSE and InfoQ.

3.1 Logistic Regression of Concentration of DNA Versus Chimerism

The concentration data was divided into two categories, one including all samples with 100% chimerism and the other including samples with chimerism less then 100%. A binary logistic regression was performed on this data with the response value defined by the chimerism category value and the concentration value as the predictor.

The result of this regression show that as the total concentration of the DNA in the mixture increases, the probability to get 100% chimerism also increases. The extended relative uncertainty of the concentration measurement is 13.17%. It is important to specify that this experiment was performed by one operator so that we were not able to analyze the reproducibility of these results.

3.2 ANOVA of Balanced One Factor Random Experiment

This analysis considers the uncertainty of the first and the second allele peak areas for a specific marker For each peak area an ANOVA test was performed in order to calculate the measurement standard deviation. The dependent variable was the peak area of the specific allele and the covariates were the placement of the allele at a specific locus, the placement of the second allele at the same locus, the DNA mixture and repetitions. The uncertainty of the chimerism of each marker was calculated according to the law of propagation. The first peak area was selected to be the peak area of the donor. The second peak area was selected to be the peak

area of the recipient. This is the same situation as when the donor and the recipient have a homozygote locus at the same marker. In that case, the second peak area of the donor and the recipient are equal to zero and the equation for chimerism result is $\frac{d}{d+r}$ Where d,r are the peak areas of the allele of the donor and the recipient, respectively. The sensitivity coefficient (partial derivative), for each allele is $\frac{\partial(chimerism)}{\partial d} = \frac{r}{(d+r)^2}$, $\frac{\partial(chimerism)}{\partial r} = \frac{d}{(d+r)^2}$.

3.3 Measurement Uncertainty Assessment of [mean%Chim]

Let, u^2_{donor} = (std. uncertainty of the donor allele)^2, $u^2_{recipient}$ = (standard uncertainty of the recipient allele)^2, c^2_{donor} = (peak area recipient mean)/(peak area recipient mean + peak area donor mean)^2. $c^2_{recipient}$ = (peak area donor mean)/(peak area recipient mean + peak area donor mean)^2. $U^2 = c^2_{donor} * u^2_{donor} + c^2_{recipient} * u^2_{recipient}$ (see Deldossi and Zappa 2009). Now that the uncertainty of each marker is known, the uncertainty of the total chimerism is calculated as:

U (total chimerism) $= \sqrt{\frac{\sum_1^n U_i^2}{n}}$ where, i is the marker index and n is the number of markers used to calculate the specific chimerism. For example, if we use all the ten markers to calculate the total chimerism, then U (total chimerism) = 0.021. For the 95% confidence level (k = 2), the total uncertainty of total chimerism is: m $\pm$ 0.042* total chimerism. If the total chimerism is 0.96, then the total chimerism will be included in the following range (0.96, 1) with the higher limit not exceeding 1. The *[mean%Chim]* is included in the following range (96,100), with the higher limit not exceeding 100%. For more on such data analysis see Koltai (2009), Chiang (2007) and Kenett and Zacks (1998).

4 Assessing PSE and Information Quality of ChimerTrack®

Practical Statistical Efficiency (PSE) consists of assessing a statistical tool or project with eight dimensions representing practical impact:

V{D}= value of the data actually collected.
V{M}= value of the statistical method employed.
V{P}= value of the problem to be solved.
V{PS}= value of the problem actually solved.
P{S}= probability that the problem actually gets solved.
P{I}= probability the solution is actually implemented.
T{I}= time the solution stays implemented.
E{R}= expected number of replications.

The overall PSE is assesses by a 1–5 scale ("1" very low, "5" very high) on each dimension and multiplying the individual scores (Kenett 2007) The PSE of the measurement uncertainty project present in this paper has been assessed as 187,500, a very high number. This reflects on the importance of this work, in particular because of the life critical aspects of the ChimerTrack® decision support system.

Information Quality (InfoQ) has been proposed as a complement to data quality and analysis quality. InfoQ is context dependent and basically considers if the right information is provided to the right decision maker, at the right time and in the right way. Information Quality consists of eight components:

- *Data resolution*
- *Data structure*
- *Data integration*
- *Temporal Relevance*
- *Generalization*
- *Chronology of Data and Goal*
- *Concept Operationalization*
- *Communication*

ChimerTrack® has a very high InfoQ potential. It provides critical data to decision makers in a timely and effective way. A more in-depth analysis of ChimerTrack® InfoQ will be presented in future work. Both PSE and InfoQ are designed to bring "the big picture" to the technical analysis (see Kenett and Shmueli 2009).

5 Summary and Conclusions

The purpose of this work is to estimate the uncertainty of the ChimerTrack® decision support system used by physicians in monitoring stem cell transplantations. It presents an example complex measurement system with several constraints and cost limitations The main steps we have covered in this case study are:

1. ***Map*** *the measurement process*.
2. ***Identify sources of data*** *that are relevant for uncertainty measurement*.
3. ***Plan*** *the statistical analysis strategy*.
4. ***Design experiments*** *to complement existing data, where needed and where possible*.
5. ***Conduct statistical analysis*** *and assess Practical Statistical Efficiency (PSE) and Information Quality* (InfoQ).
6. ***Present*** *findings*.

This six steps procedure is generic enough to be applicable in a variety of applications in medicine and beyond.

References

Applied Biosystem. (2006). AmpFlSTR® SGM Plus®PCR user's manual

Chiang A (2007). Confidence intervals for gauge repeatability and reproducibility (R&R) studies In F. Ruggeri, R. S. Kenett & F. Faltin (Eds.) *Encyclopedia of statistics in quality and reliability* Chichester, UK: Wiley

Deldossi, L. & Zappa, D. (2009). ISO5725 and GUM: comparisons and comments. *Accreditation and Quality Assurance, 3* 159–167

Kenett, R. S. (2007). Practical statistical efficiency. In F. Ruggeri, R. S. Kenett & F. Faltin, *Encyclopedia of statistics in quality and reliability* Chichester, UK: Wiley

Kenett, R. S. & Shmueli, G. (2009). *On information quality*. Submitted for publication, http://ssrn.com/abstract=1464444

Kenett, R. S., & Zacks, S. (1998). *Modern industrial statistics: Design and control of quality and reliability*. San Francisco, CA: Duxbury Press Spanish edition 2000, 2nd paperback edition 2002, Chinese edition 2004

Koltai, D. (2009). *Measurement uncertainty in quantitative chimerism: monitoring after stem cell transplantation*. MSc thesis, Bar Ilan University, Israel

Kristt, D., & Klein, T. (2004). STR-based chimerism testing: using ChimerTrack® interactive-graphics software to ease the burden. *ASHI Quarterly, 28* 16–19

Kristt, D., Stein, J., Yaniv, I., & Klein, T. (2004). Interactive ChimerTrack® software facilitates computation, visual displays and long-term tracking of chimeric status based on STRs. *Leukemia, 18*(5),1–3

Kristt, D., Israeli, M., Narinski, R., Or, H., Yaniv, I., Stein, J., et al. (2005). Hematopoietic chimerism monitoring based on STRs: quantitative platform performance on sequential samples. *Journal of Biomolecular Techniques, 16* 1–28

Kristt, D., Stein, J., Yaniv, I., & Klein, T. (2007). Assessing quantitative chimerism longitudinally: technical considerations, clinical applications and routine feasibility. *Bone Marrow Transplantation, 39* 255–268

Satisfaction, Loyalty and WOM in Dental Care Sector

Paolo Mariani and Emma Zavarrone

Abstract We propose two different measures of the dental care industry: (a) BALG matrix, a new instrument to measure patient loyalty and its extent; (b) SERVQUAL based approach to measure patient satisfaction. Further investigation concerns the link between patient satisfaction and loyalty. The results prove that patient loyalty in the dental care industry is similar to consumer behaviour in all the other B2B and B2C services and furthermore, the results highlight low dependency of patient satisfaction on loyalty.

1 Introduction

Concepts such as satisfaction, loyalty, acquisition and retention represent CRM (Customer Relationship Management) key strategies which are becoming increasingly common within B2B (Business to Business) and B2C (Business to Consumer) markets; and their comprehension, measurement and aware management are key factors of success (Oliveira Lima 2009). Nonetheless, the processes of activating retention and acquisition are sometimes impossible as either the competencies or the financial efforts they both require can be unaffordable. Retention, for example intended as keeping communication open with customers can be difficult to achieve as it requires effective programming to receive and respond to complaints, to develop long-term relationships by meeting their changing needs. An acquisition strategy implies monitoring of different acquisition channels (direct mail, telephone solicitation, etc.). The creation of retention and acquisition programs in the dental care industry is under researched and made difficult by the characteristics of the service itself and by effective fear of the dentistry. For these reasons, we only focus our research on the strategic tools for measuring: patient satisfaction and loyalty.

The remainder of the paper is organized as follows: Sect. 2 deals with loyalty concept and content; Sect. 3 examines patient satisfaction; Sect. 4 illustrates the hypotheses of our approach and the related measures; Sect. 5 presents a first application of the suggested methodology.

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_19,

2 Patient Loyalty

Loyalty is not a straightforward construct and academics are still searching for the most suitable approaches (Bolton et al. 2000). Over the past 30 years, academics have debated loyalty and its core issue, dimension and measurement of the concept. Jacoby and Chestnut (1978) explored the psychological meaning of loyalty, drawing attention to the idea that repeated purchase by itself was not a reliable signal of loyalty, but just a component. Adopting the same approach Gremler and Brown (1996) conceive loyalty as consisting of at least four dimensions, which the authors synthesise into the following definition: loyalty is "the degree to which a customer exhibits repeat purchasing behavior from a service provider, possesses a positive attitudinal disposition toward the provider, and considers using only this provider when a need for this service arises" which introduces to the framework developed by Oliver (1999) which introduces the framework developed by Oliver (1999) which extends the notion of incorporating repeated purchase and divides loyalty into three dimensions: cognitive, affective or attitudinal and conative or behavioral. The identified markers for the three dimensions are: price and service features for cognitive loyalty; repeated purchase for behavioral loyalty and positive word of mouth (WOM) for attitudinal loyalty. In detail, positive word-of-mouth is a common approach to loyalty conceptualization. Loyal customers become an advocate for the service (Payne 1993). Following Butcher et al. (2001) positive WOM can be analyzed from different points of view:

- Providing positive word-of-mouth (Zeithaml et al. 1996, Andreassen and Lindestad 1998).
- Recommending the service to others (Stum and Thiry 1991, Fisk et al. 1990).
- Encouraging others to use the service (Kingstrom 1983, Bettencourt and Brown 1997).
- Defending the service provider's virtues (Kingstrom 1983).

Once attitudinal and behavioral loyalty measures are gathered researcher may respect the dimensionality of loyalty treating the two variables as separate constructs and using them to categorise the loyalty forms proposed by Baldinger and Rubinson (1996) and shown in Fig. 1. While a satisfied customer is merely a passive recipient of service the loyal customer feels a positive connection to the service provider according McGarry (1995), thus loyal customers become active ambassadors for the business (Butcher et al. 2001).

3 Patient Satisfaction

The concept of patient satisfaction cannot be considered separately from that of service quality, as demonstrated by Iacobucci (1995). However, it has been proved that patients' perceptions frequently differ from those of doctors, and that doctors frequently misperceive their patients' perceptions (Brown and Payne 1986). Thus, the provision of dental care which is technically correct from a dentist's point of

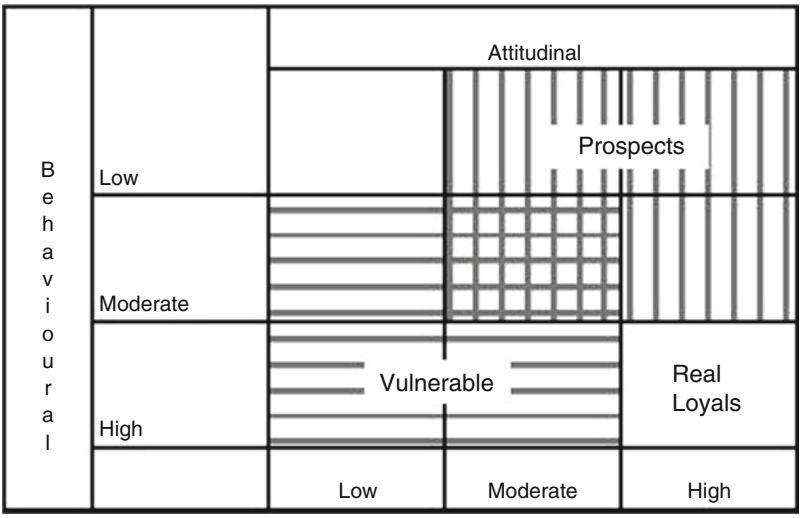


Fig. 1 Behavioral attitudinal loyalty matrix

view, but is provided in a manner which is less than desirable to the patient and thus may provoke a low patient evaluation(s) of the service. This could affect patient satisfaction and, consequently loyalty. The causes which drive satisfaction have been widely studied in literature for the last 20 years; several and very complex aspects have been examined from correctly identifying the concept to measuring it.

Satisfaction can certainly be defined a cross-sectional dimension and it plays a central role in the study of customer segmentation. The concept of satisfaction analysed with reference to patients can be synthesised by the definition of Keegan et al. (1989): "... patient satisfaction is an attitude – a person's general orientation towards a total experience of health care. Satisfaction comprises both cognitive and emotional facets and relates to previous experiences, expectations and social networks". According to authors, cognitive and emotional aspects should be made operative and compared to perceptions, as proposed by Oliver's contribution (1999), well known as the disconfirmation paradigm. This paradigm is still up-to-date, although several measurement difficulties exist, especially with reference to quantifying expectations. For this reason, the main methods proposed rely on appositely created scales which aim at overcoming the limit of expectations. The present work follows this approach as well, therefore expectations are not directly measured, even though they are linked to patients' evaluations on the different aspects of the service.

4 Hypothesis and Measures

4.1 Hypothesis

The model moves from considering that satisfaction and loyalty, have to be investigated from a subjective perspective given their nature In detail we hypothesize

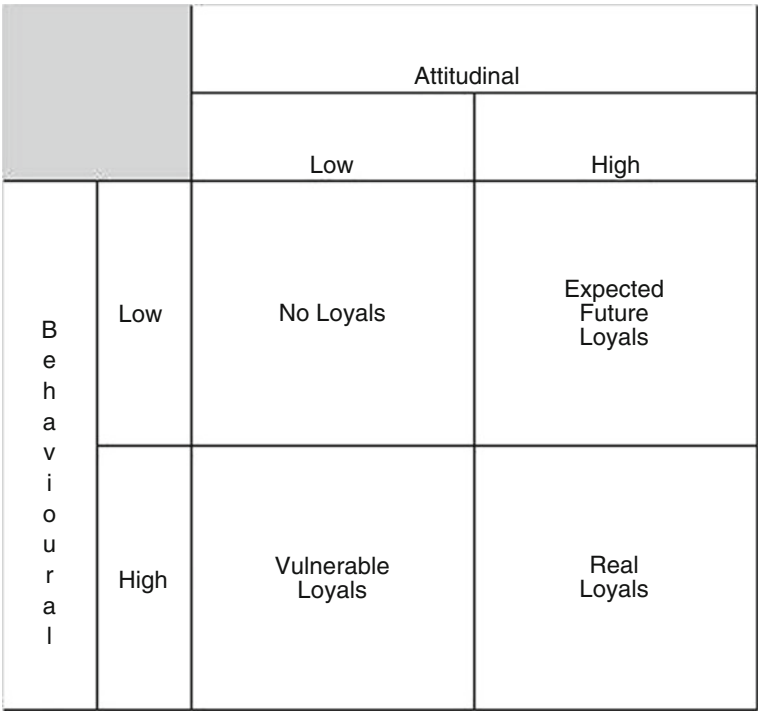


Fig. 2 BALG matrix

that a relationship exists between attitudinal and behavioral loyalty in the dental care industry and this can be used to create customer strategies' for each group of patients. Thus, we disentangle patient loyalty in the dental care industry from the first two definitions and we only consider repeated purchase as a proxy of behavioral loyalty and positive WOM and referral as an attitudinal aspect. Thus, the proposed tool following the indication of Baldinger and Rubinson (1996) reduces their dimensions in four clusters. This matrix is defined as BALG (Fig. 2).

4.2 Measures

The measures proposed in this work derive from the following three steps:

- We operationalize the attitudinal and behavioral concept;
- We deal with quantifying patient satisfaction through a new approach;
- We use these new measures to analyse the relationship between satisfaction and loyalty.

In the first step we select only the repeated purchase and translate it in the dental industry as the return as proxy of behavioral loyalty (B). This measure is

dichotomised as 1 to indicate patients who returned after the first year and 0 the patients who returned during the year. For the attitudinal loyalty we suggest the use of two items: firstly: "How Did You Learn About This Dental Office?" and the answer is codified in a multiple choice leaves from family to internet; the second one is "Would you recommend this dental office to other individuals?" with dichotomous answers. We then construct a new variable based on these two items in which we can measure the positive WOM with family referral (code 1) and the negative WOM with generic referral (code 0). Then we construct the BALG matrix from the two variables.

In the second step, we use a modified approach based on SERVQUAL to measure patient satisfaction. The measure proposed for patient satisfaction recalls the SERVQUAL scheme for the gaps mode only. The measurement components of the model differ as we include the quantification of the degree of agreement among n patients about m service aspects, their importance and their weight, obviously, provided by the dentist's judgment. According to this approach, patient satisfaction (PS) derives from two components: overall evaluation of service delivered (PE) and the importance of the service delivered (IP) measured on n patients for the distinct m dimensions of the service. This gap is then corrected for a system of weights (W), representing the dentist's perspective. This system is obtained through a Delphi analysis. In this framework, we can define the patient satisfaction index as follows:

$$PSI = \sum_{j=1}^{m} D_j W_j \tag{1}$$

$$D = \sum_{i=1}^{n} (PE_i - IP_i) \tag{2}$$

PE_i: overall judgment on service delivered;

IP_i: importance of service delivered.

To verify the hypothesis, we proceeded as follows: identification of the patient satisfaction aspects (focus group on eight patients and five doctors) patient satisfaction aspects formulation of a questionnaire covering the various dimensions (Tangible Elements, Reception, Dental Treatment, Dental Hygienist, Dental Assistant, Dentist Surgeon); questionnaire pre-test (analysis on validity and reliability with test and retest method); reformulation of some items; formulation the final version of the questionnaire. Questions were formulated following the gap scheme, i.e., each question asked provided two different scores on a scale (from 1 to 6) regarding either different aspects of the dimension investigated or its importance. The result of this long process provides the patient satisfaction index (PSI).

5 Application and Main Findings

A random sample of 104 patients was selected in a dental care centre in Milan during the first quarter 2007. They were provided with the ad hoc questionnaire which was returned by 84% of patients, but some questionnaire presented partial missing.

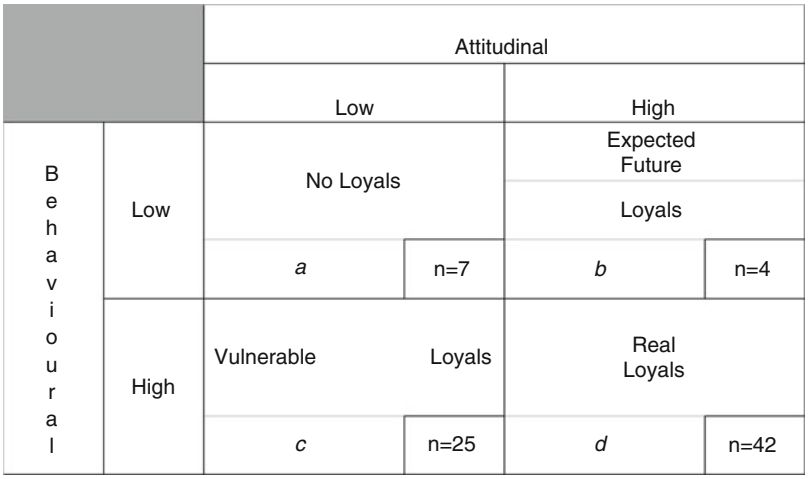


Fig. 3 BALG size

41% of the respondents had been a client for less than a year with an average age ranging from 25 to 59. The analysis can be divided in two parts: the first one regards the application of BALG matrix while the second one shows the PSI and the relationship between loyalty and patient satisfaction. The BALG matrix is shown in Fig. 3, with the size of each clusters. To obtain a measure of association from Table 1, we use the well-known Phi (SPSS 16.0 User's Guide 2007), a measure of the degree of association between two binary variables. This measure is similar to the correlation coefficient in its interpretation:

$$Phi = \frac{ad - bc}{\sqrt{(a+b) * (a+c) * (c+d) * (b+d)}} \tag{3}$$

The BALG matrix highlights a low positive association (*Phi* coefficient equals 0.186). This effect is masked by age. In fact, we compute the BALG matrix for each quarter of age and thus we argue that association is more intense for adult and older patients whose *Phi* value is equal to 0.231 and 0.344, respectively. This result shows that patient loyalty increases with experienced of special care. Moreover, we can conclude that BALG increases as patient age increases. The second part of the study regards the analysis of the proxy of satisfaction that outlines very high average scores. The means for all the six sections of the questionnaire respectively, are 5.93, 5.97, 6.00, 5.97, 5.98, 6.00 with very low values of the coefficient of variation. Overall, the comparison between the BALG matrix clusters and PSI score (Fig. 4) shows high variability as confirmed by the normalized coefficient of variation (denoted by cv^*). To verify a possible association between patient satisfaction and attitudinal and behavioural loyalty we dichotomised the above mentioned patient satisfaction score for each six aspects after comparison with a single area of the BALG matrix

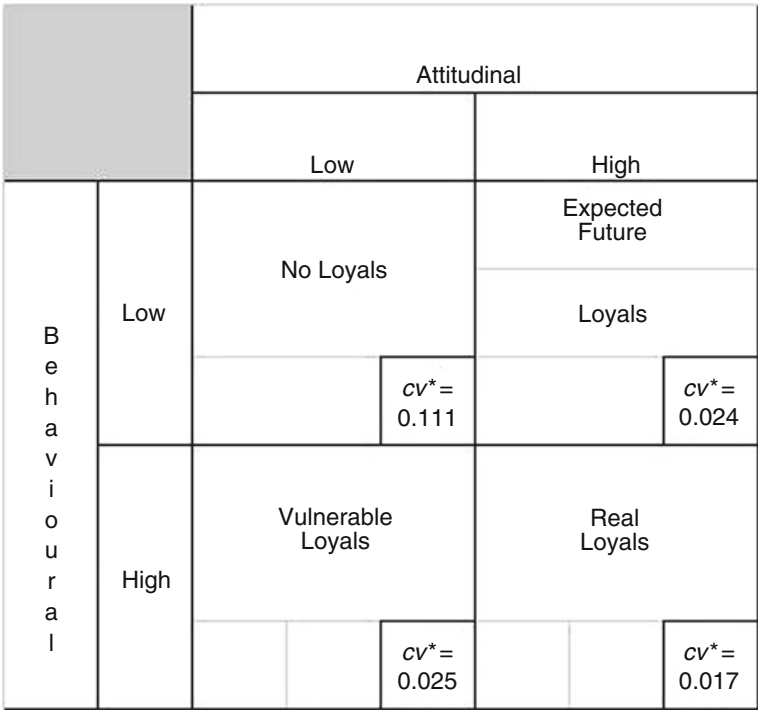


Fig. 4 BALG and PSI scores

Table 1 PS on reception and behavioral loyalty

		PS reception		Total
		Low	High	
B	Low	7	4	11
	High	20	47	67
Total		67	21	78

(*Phi* = 0.247 sign. 0.029) (Table 1). Thus, a precise result concerning the nature of the relationship between loyalty and satisfaction cannot be provided as it depends on the service aspects like any other industry.

Acknowledgements We thank Drs. Elio Marino and John Kois for permission to use their data.

References

Andreassen, T. W., & Lindestad, B. (1998). Customer loyalty and complex services: The impact of corporate image on quality, customer satisfaction and loyalty for customers with varying degrees of service expertise. *International Journal of Service Industry Management*, *9*, 7–23.

Baldinger, A., & Rubinson, J. (1996). Brand loyalty: The link between attitude and behaviour. *Journal of Adverting Research*, *36*(6), 22–34.

Bettencourt, L. A., & Brown, S. W. (1997). Contact employees: Relationships among workplace fairness, job satisfaction and prosocial service behaviors. *Journal of Retailing*, *73*(1), 39–61.

Bolton, R., et al. (2000). Implications of loyalty program membership and service experiences for customer retention and value. *Journal of the Academy of Marketing Science*, *28*(1), 95–108.

Brown, S., & Swartz, T. (1989). A gap analysis of professional service quality. *The Journal of Marketing*, *53*, 92–98.

Butcher, K., Sparks, B., & O'Callaghan, F. (2001). Evalutative and relational influences on service loyalty. *International Journal of Service Industry Management*, *12*(4), 310–327.

Cronin, J. J., & Taylor, S. A. (1994). SERVPERF versus SERVQUAL: Reconciling performance-based and perceptions-minus-expectations measurement of service quality. *Journal of Marketing*, *58*(1), 125–132.

De Oliveira Lima E. (2009). *Domain knowledge integration in data mining for churn and customer lifetime value modeling: New approach and applications*, University of Southampton (UK), School of Management, PhD Thesis, 12–26, http://eprints.soton.ac.uk/65692/

Fisk, T. A., Brown, C. J., Cannizzaro, K. G., & Naftal, B. (1990). Creating patient satisfaction and loyalty. *Journal of Health Care Marketing*, *10*(2), 5–15.

Gremler, D. D., & Brown, S. W. (1996). Service loyalty: Its nature, importance, and implications. In B. Edvardsson, et al. (Eds.), *Advancing service quality: A global perspective*, International Service Quality Association pp. 171–80.

Hayes, B. E. (2008). *Measuring customer satisfaction and loyalty, survey design, use, and statistical analysis method* (3rd ed.). Milwaukee, Wisconsin: ASQ Quality Press.

Iacobucci, D. (1995). Distinguishing service quality and customer satisfaction: The voice of the consumer. *Journal of Consumer Psychology*, *4*(3), 277–303.

Jacoby, J., & Chestnut, R. W. (1978). *Brand loyalty: Measurement and management*. New York: John Wiley and Sons.

Keegan, D. P., Eiler, R. G., & Jones, C. R. (1989). Are your performance measures obsolete? *Management Accounting*, *70*(12), 45–50.

Kingstrom, P. O. (1983). Patient ties to ambulatory care providers: The concept of provider Loyalty. *Journal of Health Care Marketing*, *3*(2), 27–34.

McGarry, D. (1995). The road to customer loyalty. *Canadian Business Review*, *22*(1), 35–61.

Oliver, R. (1999). Whence consumer loyalty? *Journal of Marketing Research*, *63*, 33–44.

Payne, A. (1993). *The essence of services marketing*. London: Prentice-Hall.

Steffes, M., Murthi, E., Rao, B., & Ram, C. (2008). Acquisition, affinity and rewards: Do they stay or do they go? *Journal of Financial Services Marketing*, *13*(3), 221–233.

Stum, D. L., & Thiry, A. (1991). Building customer loyalty. *Training and Development Journal*, *45*, 34–36.

SPSS 16.0 User's Guide (2007). Guide to Data Analysis, Prentice Hall, Upper Saddle River, New Jersey, USA.

Zeithaml, V. A., Berry, L. L., & Parasuraman, A. (1996). The behavioral consequences of service quality. *Journal of Marketing*, *63*, 33–44.

Controlled Calibration in Presence of Clustered Measures

Silvia Salini and Nadia Solaro

Abstract In the context of statistical controlled calibration we introduce the 'multilevel calibration estimator' in order to account for clustered measurements. To tackle this issue more closely, results from a simulation study will be extensively discussed. Finally, an application from a building industry will be presented.

1 Introduction

Statistical calibration is a set of procedures developed to recover a true, unknown measure from a single measure or a plurality of approximated measures available for the same item and derived from alternative measuring instruments. The literature on statistical calibration is extremely rich. Many different approaches have been developed over the years. A comprehensive monograph is given by Brown (1993), whilst an excellent review on these methods along with the debate that throughout had surrounding them is contained in Osborne (1991).

The most popular approach, though not without criticism (Brown 1993; Osborne 1991), is the classical calibration, perhaps for its simplicity. Let X and Y be two different measurements for the same item, the first more accurate but also more expensive to obtain than the second. Classical calibration relies on the two following steps. At calibration step, a so-called training set is formed by randomly drawing n items on which measures (x_i, y_i) are taken, $(i = 1, \ldots, n)$. This is the calibration experiment. In particular, in controlled calibration the x-values are kept fixed during the experiment. Once these measures are obtained, the calibration model expressing the relationship linking Y with X can be estimated. Usually, a linear model is involved: $Y_i = b_0 + b_1 x_i + \varepsilon_i$, for all i, where the model errors ε_i are assumed i.i.d. $N(0, \sigma^2)$. Parameter estimates $\hat{b}_0$ and $\hat{b}_1$ are then obtained through the OLS method. At prediction step, where y-values only are collected, the values of X will be predicted by relying on the model estimated during the calibration step. Formally, let Y_0 be a new observation on Y and ξ the unknown value of X. The classical calibration estimator $\hat{\xi}_C$ for ξ is: $\hat{\xi}_C = \frac{Y_0 - \hat{b}_0}{\hat{b}_1}$, (see Brown 1993). Under normal errors

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_20,

ε_i, $\hat{\xi}_C$ is the maximum likelihood (ML) estimator of ξ. Asymptotic properties derive then straightforwardly.

Recently, certain authors have shown that in presence of outliers or groups in data the classical estimator generally fails to match the expected performance, (e.g. yielding inefficient estimates). Several methods have been proposed to solve the problem regarding the influence of outliers (e.g. Cheng and Van Ness 1997). By dealing with data that include either groups or just the outliers, Salini (2006) compared the performance of the classical estimator with the two estimators derived by replacing in its formula, respectively, Huber's and Tukey's robust estimates for b_0 and b_1. Even if robust estimators perform better in the presence of outliers, it has been shown that their performance is similar to the classical estimator if measures are clustered.

2 The Multilevel Calibration Estimator and its Forerunner

In order to tackle the problem of clustered measures we propose to use a multilevel model at calibration step, thus allowing within- and between-groups linear relations linking Y with X to be represented in a unified formulation. At prediction step, we introduce the 'multilevel calibration estimator' (MLVCE), so defined in order to include the cluster information of a new item in a simple, straight way. Assume that data follow a two-level hierarchy, where J is the number of groups or clusters (level-2 units) each including n_j items (level-1 units), ($j = 1, \ldots, J$). Let Y and X be two measurements jointly taken on level-1 units. Given that here two measurements only are involved (univariate calibration), we can simply rely on a random–slopes model:

$$Y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij}, \qquad i = 1, \ldots, n_j; j = 1, \ldots, J, \tag{1}$$

where the level-1 errors ε_{ij} are i.i.d. $N(0, \sigma^2)$, and β_{0j} and β_{1j} are the random parameters expressing different between-groups linear relationships of Y and X. Such parameters can be explicitly modeled by level-2 relations. In this work we are going to consider:

$$\beta_{0j} = \gamma_{00} + U_{0j} \qquad \text{and:} \qquad \beta_{1j} = \gamma_{10} + U_{1j}, \tag{2}$$

for all j. Another example is: $\beta_{0j} = \gamma_{00} + \gamma_{01} w_j + U_{0j}$ and: $\beta_{1j} = \gamma_{10} + \gamma_{11} w_j + U_{1j}$, where W denotes a possibly available level-2 measurement, i.e. a measurement regarding the clusters and not directly the items within clusters. Parameters γ appearing in formulas (2) are said fixed effects, while the Us are called random effects. Usually, vectors $\mathbf{U}_j = (U_{0j}, U_{1j})'$ are i.i.d. $N_2(\mathbf{0}, \mathbf{T})$, for all j, and are independent of level-1 errors. As known, all the parameters in the model (1)–(2) can be estimated through the ML approach. For the variance–covariance components σ^2 and τ_{lm} in matrix $\mathbf{T}$, ($l, m = 0, 1$), full ML or restricted ML estimators (REML) can be derived (see e.g. Pinheiro and Bates 2000). Finally, the best linear

unbiased predictors of the random effects U are obtained as empirical Bayes estimators (see Pinheiro and Bates 2000). For further details see Goldstein (1995) and Snijders and Bosker (1999).

At prediction step, the main concern is to predict the value ξ of X of a new additional item, not included in the initial dataset, by using its cluster information and knowing its measure Y. We denote this new observation with Y_{0j^*} in order to stress the fact that it derives from group j^*, which is one of the J observed groups. Then, to predict the corresponding value ξ_{j^*} we rely on the MLVCE, here indicated with $\hat{\xi}^m_{j^*}$. Under the relations (2), and conditionally to the calibration step, it is given by:

$$\hat{\xi}^m_{j^*} = \frac{Y_{0j^*} - \hat{\gamma}_{00} - \hat{u}_{0j^*}}{\hat{\gamma}_{10} + \hat{u}_{1j^*}}, \tag{3}$$

where $\hat{u}$ in (3) indicates the empirical Bayes prediction for the random effect U in the group j^*. From a theoretical point of view, the estimator (3) could be further generalized to embody more complex level-2 relations than (2). We will not however dissect such topic here. It will be more exhaustively treated elsewhere.

The proposal we have just described draws its basic idea from Oman's approach for calibrating in the presence of repeated measurements (Oman 1998), the level-2 units consisting of individuals to be measured and the level-1 units of measures taken repeatedly on them. Being repeated measures an issue regarding clustered data, the statistical tools employed here are very similar to Oman's. In any case, our proposal differs from Oman's in its essence at both steps. At calibration step, Oman actually introduces a random–slopes model of the same form given in (1) and under the same assumptions. However, by formulating the problem without making explicit reference to level-2 relations, he excludes, in principle, the use of more complex level-2 relations than (2) and then the possibility to introduce level-2 measurements to explain between-groups differences. At prediction step, Oman assumes that a measure Y_0 is available for a new individual. This is like having a new cluster, instead of a new item that just belongs to a given cluster. Oman then defines his classical estimator (OCE) for the value ξ of this new individual conditionally to the ML fixed-effects estimates $\hat{\gamma}_{00}$ and $\hat{\gamma}_{10}$ computed at the calibration step: $\hat{\xi}^{om}_C = \frac{Y_0 - \hat{\gamma}_{00}}{\hat{\gamma}_{10}}$. Not disposing of many additional information, his prediction is therefore based on population-averaged parameters. Oman admits that the performance of the OCE can be improved in terms of mean squared error. Thus, he brings in two further variations. The first is the contraction estimator, derived by minimizing $\mathrm{E}(c\hat{\xi}^{om}_C - \xi)^2$ over c. The second is the ML estimator of ξ obtained from: $\hat{\xi}^{om}_C \sim N(\xi, \mathrm{Var}(\hat{\xi}^{om}_C))$, given $\hat{\gamma}_{00}$ and $\hat{\gamma}_{10}$. For all technical details see Sect. 3 by Oman (1998).

Further remarks about the comparison between our approach and Oman's should be made. Firstly, in the form here proposed the MLVCE is based on a double conditioning procedure. In addition to the calibration step, there is also a conditioning on the knowledge of the group. Hence, the variance of the MLVCE would involve random-effect predictions as given terms. This point is not trivial. To account for the whole uncertainty occurring at the prediction step, (i.e. the cluster membership of the item that is about to appear), one should consider that formula (3) contains

the ratio of two random variables, whose variance cannot be derived in a closed form. This problem, however, is beyond the scope of this current study, where the primary concern is to start tackling the problem through explorative ways. Secondly, the MLVCE (3) could be also viewed as an improvement of the performance of the OCE when the group membership of the new item is known. In his aforementioned work, Oman actually introduces a further variant of the OCE to include additional information about Y_0, if available (Oman 1998, p. 443). Viewed in our framework, it looks like a method to embody the group membership in the prediction procedure. However, our approach is substantially different. All the information available on the group j^* are used at the calibration step to produce the parameter estimates of the model (1) and (2), which are subsequently used to compute the multilevel estimate (3). In Oman instead, being the additional information possibly available only when a new individual appears, these are incorporated in the estimation process at the prediction step, thus adjusting Oman's classical estimate *a posteriori*.

3 The Simulation Study

In order to give a first insight into the main statistical properties of the MLVCE (3), we performed an extensive simulation study involving the model (1) and (2) under various sets of experimental conditions. We were also interested in comparing the performance of the MLVCE with the other estimators, namely classical, Oman's classical, contraction and ML estimators. In addition, we introduce the 'within-group classical calibration estimator' (WGCE), simply given by: $\hat{\xi}_{C_j} = \frac{Y_0 - \hat{b}_{0j}}{\hat{b}_{1j}}$, where $\hat{b}_{0j}$ and $\hat{b}_{1j}$ are the OLS estimates for parameters b_{0j} and b_{1j} of the J linear regressions separately fitted within each group, ($j = 1, \ldots, J$).

Our empirical experience on calibration data prompted us to keep two main features directly in control, i.e. within-group correlation of X and Y along with variances and covariances of random effects. While the first issue is deeply–rooted in statistical calibration, representing how reliable the measure Y is for X, the second directly stems from the MLVCE including random–effects predictions. Therefore, to yield data that could reflect the above aspects, we obtained pseudo-observations y^*_{ij} forming variable Y^* as balanced data (i.e. $n_j = n, \forall j$) through equation (1) re-expressed as: $y^*_{ij} = \beta^*_{0j} + \beta^*_{1j} x^*_{ij} + \omega_j \varepsilon^*_{ij}$, for all i, j, where ω_j is a strictly positive weight introduced to guarantee a desired level of correlation of X and Y within group j and which requires to be determined accordingly, as it will be shown soon. Firstly, variables X^* and ε^* were formed, respectively, by a priori assigning nJ values x^*_{ij}, with common mean and variance ($\sigma^2_{X^*}$) over groups, and by randomly drawing nJ values ε^*_{ij} from a $N(0, \sigma^2_{\varepsilon^*})$, $\sigma^2_{\varepsilon^*}$ fixed and common over groups. Parameters β^*_{0j} and β^*_{1j} were computed from level-2 equations (2) after generating J random-effects values u^*_{0j} and u^*_{1j} from a $N_2(\mathbf{0}, \mathbf{T})$, $\mathbf{T}$ fixed, and with γ_{00} and γ_{10} arbitrarily chosen. Then, let Y^*_j and X^*_j be the set of values of Y^* and X^*, respectively, pertaining to group j. Given that in each group it

Table 1 Simulation study: target values and experimental conditions

• *Target values*	Fixed effects: $\gamma_{00} = 15$, $\gamma_{10} = 3$; Residual variance: $\sigma^2_{\varepsilon^*} = 20$; Variable X: nJ values x^*_{ij} from $N(30, 40)$ and kept fixed during simulations
• *Experimental conditions*	(with fixed $J = 10, n = 100$)
1st analysis	REV (Random-Effects Var.) & WCC (Within-group Corr. Coeff.); $\rho_{01} = 0$
REV:	(a) $\tau_{00} = 10$, $\tau_{11} = 10$; (b) $\tau_{00} = 10$, $\tau_{11} = 5$; (c) $\tau_{00} = 10$, $\tau_{11} = 1$; (d) $\tau_{00} = 100$, $\tau_{11} = 1$
WCC:	(A) $\rho^* = 0.80$; (B) $\rho^* = 0.85$; (C) $\rho^* = 0.90$; (D) $\rho^* = 0.95$
2nd analysis	REV & RECC (Random-Effects Correlation Coefficient); $\rho^* = 0.85$
REV:	cases (a)–(d) RECC: (I) $\rho_{01} = 0.4$; (II) $\rho_{01} = 0.9$

holds: $\mathrm{Var}(Y^*_j) = \sigma^2_{Y^*_j} = \beta^{2*}_{1j}\sigma^2_{X^*} + \omega^2_j\sigma^2_{\varepsilon^*}$, and: $\mathrm{Cov}(X^*_j, Y^*_j) = \beta^*_{1j}\sigma^2_{X^*}$, the correlation coefficient ρ_j of X^*_j and Y^*_j is given by: $\rho_j = \beta^*_{1j}\sigma_{X^*}/\sigma_{Y^*_j}$, from which straight computations lead to the following formula regarding the weight: $\omega_j = \sqrt{(1-\rho^2_j)/\rho^2_j}\,\frac{\sigma_{X^*}}{\sigma_{\varepsilon^*}}\,\mathrm{abs}\,(\beta^*_{1j})$, for all j, where the parameters ρ_j, $\sigma^2_{X^*}$ and $\sigma^2_{\varepsilon^*}$ can be fixed at a desired level. As it is apparent, weights may introduce a certain degree of heteroskedasticity at level-1. Although this might be considered as potentially undesirable, working with different ω_j brings the advantage, at the same time, of keeping ρ_j fixed to a nominal value ρ^* equal for all groups, and allowing the random–effects variance–covariance matrix $\mathbf{T}$ to vary.

At this point, to reach our objectives we divided the simulation study into two distinct kinds of analyses. Table 1 resumes the experimental conditions along with the target values we assumed without loss of generality. The first analysis is addressed at evaluating the influence of random–effects variances in connection with within-group correlation of X and Y. Random effects are here assumed to be uncorrelated. In our experience, the performance of the MLVCE, if viewed in terms of prediction error, may depend on the magnitude of the random slope variance (τ_{11}), or even on its ratio with the random intercept variance (τ_{00}). This motivated our choice regarding the variance values appearing in Table 1 under the acronym REV. In particular, we considered four different couples of values (from (a) to (d), Table 1) such that the ratio of variances $\frac{\tau_{11}}{\tau_{00}}$ is: $1, 0.5, 0.1, 0.01$, respectively. As for the within-group correlation coefficient of X and Y (WCC), four different levels were considered (from moderately high (A) to very high (D), Table 1). We confined our attention to strongly linear relationships representing the implicit, fundamental condition for resorting to a calibration approach. The second analysis was conceived to study in depth the role of the random–effects variance–covariance structure. In addition to the four cases (a)–(d) concerning variances, a positive correlation of the random effects (ρ_{01}) was introduced (intermediate level (I) and high level (II), Table 1), with the WCC kept fixed. Both analyses were carried out by constantly keeping the group size n and the group number J fixed to: $J = 10$ and $n = 100$, in that this represents one of the situations we had frequently encountered in calibration problems. Nonetheless, we are going to undertake more in-depth inspections about

the sample size role in a future work. Finally, $K = 1{,}000$ simulation runs were carried out for each combination of the experimental conditions. We implemented all the simulation routines in R, vs. 2.9.0. Parameters in the model (1) and (2) were estimated through full ML and REML procedures in the library 'nlme' by Pinheiro et al. (2011). Next, we evaluated the performance of the six estimators in terms of both estimation error and prediction error. Being the prediction error more in line with calibration intents, the attention is here confined only to it. Prediction error was evaluated through the Root Mean Squared Error of Prediction: $\text{RMSEP} = \sqrt{\frac{\text{PRESS}}{df\,K}} = \sqrt{\frac{\sum_{k=1}^{K}\sum_{l=1}^{N}(\hat{\xi}_{k(l)}-\xi_l)^2}{df\,K}}$, where PRESS, i.e. the predicted-residual-sum-of-squares statistic, is determined through a cross-validation process with leave-one-out jackknife (Shiao and Tu 1995). The RMSEP is adjusted for the degrees of freedom (df), they being different depending on the estimator type. In particular: $df = N - 1 - C_{\text{estim.}}$, where $N = nJ$ and $C_{\text{estim.}}$ is given by: $C_{\text{class.}} = C_{\text{oman}} = 2$ for classical and Oman's estimators; $C_{\text{within-gr.}} = 2J$ for the WGCE; $C_{\text{multilev.}} = J + 1$ for the MLVCE, according to the method for counting degrees of freedom adopted by Pinheiro and Bates (2000, p. 91).

Simulation Results Before discussing simulation results, two remarks should be made. Firstly, a certain number of non-convergences occurred during the cross-validation process when estimating the parameters in model (1) and (2). The main problem was the singularity of the hessian matrix in the optimization of the log-likelihood function at the calibration step. The RMSEP formula was then modified accordingly, by substituting: $df\,K$ with: $\sum_k^K df_k = \sum_k^K (c_k - C_{\text{estim.}})$, where c_k expresses the number of convergences at the k-th run ($c_k \leq N - 1,\ \forall k$). Secondly, a stability analysis of cross-validation results made through the monitoring of the statistic RMSEP_k, i.e. the RMSEP computed up to the k-th run for all k, revealed that in all estimators a small amount of outlying estimates were present. This may depend on potential anomalies generated in the data. Then, in order to avoid wrong conclusions we decided to discard extreme PRESS values. Under the same experimental conditions and for each estimator, we considered a PRESS deviation p as extreme if it is outside the interval $[p_{0.25} - 3\text{IQR}, p_{0.75} + 3\text{IQR}]$, with: $\text{IQR} = p_{0.75} - p_{0.25}$, i.e. the interquartile range of PRESS deviations. Having never detected an 'extremely small' p, only values greater than $p_{0.75} + 3\text{IQR}$ were removed.

Part A of Table 2 reports the RMSEP results computed for the MLVCE (3) for both the analyses, after removing extreme PRESS values. As expected, in the first analysis (left-hand table) we note that for each REV case the RMSEP tends to reduce monotonically, while the WCC increases. A similar, if a bit slighter tendency, can be observed for each WCC as well, when the ratio $\frac{\tau_{11}}{\tau_{00}}$ decreases from 1 to 0.01. The second analysis (right-hand table) shares this last feature as well. However, the effect of the random–effects correlation coefficient (RECC) is not so evident.

The WGCE, whose results are omitted here because of the limited space, attains RMSEP values that are constantly slightly smaller than the MLVCE. However, such a difference appears to cancel out as the ratio $\frac{\tau_{11}}{\tau_{00}}$ decreases from 1 to 0.01. This

Table 2 RMSEP results for the various calibration estimators

	1ST ANALYSIS: WCC				2ND ANALYSIS: RECC		
	0.80	0.85	0.90	0.95	0	0.4	0.9
*Part A: Multilevel calibration estimator (*REML*)*							
REV case (a)	4.394	3.636	2.851	1.940	3.636	3.625	3.623
REV case (b)	4.386	3.623	2.841	1.934	3.623	3.622	3.612
REV case (c)	4.335	3.596	2.810	1.908	3.596	3.600	3.610
REV case (d)	4.334	3.590	2.804	1.908	3.590	3.586	3.590
Part B: The other estimators in REV case (d)							
Classical	9.970	9.694	9.370	9.225	9.694	10.534	11.833
Oman's classical	9.936	9.648	9.280	9.183	9.648	10.605	11.768
Oman's contraction	9.733	9.488	9.031	8.901	9.488	10.397	11.657
Oman's ML	9.541	9.290	8.872	8.736	9.290	10.124	11.246

seems to us a very interesting result, because the MLVCE turned out to perform similarly to the within-group estimator, even though with $J - 1$ parameters less.

Part B of Table 2 displays RMSEP results for the classical and Oman's estimators, respectively, when they are seen in one of their 'best' performances, which corresponds to the REV case (d). Several points are worth noting. In both analyses Oman's ML estimator attains the smallest RMSEP values under all the experimental conditions. Furthermore, in the first analysis all the estimators exhibit the same tendency highlighted in the multilevel case, so that the RMSEP tends to decrease as the WCC increases. As for the second analysis, the RMSEP tends to increase with the increasing of the RECC. This is a new aspect, occurring in neither the MLVCE nor the within-group estimator. Results for the other REV cases (a)–(c) are here omitted, the trends being similar to Table 2. The only difference lies in the prediction error magnitude, which is for both analyses approximately in the order of 17–18 in REV case (a), 15–16 in case (b) and very similar to Table 2, Part B, in case (c).

Finally, while the MLVCE has shown a satisfactory performance, it has proved to be particularly sensitive to the presence of very anomalous groups. Since these (few) groups seem mostly defined by a much lower variability of Y than the others in the same dataset, we figure they were engendered by the weights ω_j adopted for data generation. Even though we argue that such situations are far from being representative of real calibration settings, they would undoubtedly deserve more in-depth inspections than those practicable in this present context.

4 Application: Concrete Blocks Data

Concrete is a construction material composed of cement along with other cementitious materials, such as fly ash and slag cement, coarse aggregates, water, and chemical admixtures. One of the main properties of the concrete is strength. In general, its measuring gives rise to a calibration problem, since the standard method through which determining the exact grade of strength X is destructive. The alternative

method giving measures Y consists of using a sclerometer, a non-destructive instrument for empirical tests of resistance to pressure.

As a brief illustration, we present the analyses carried out on concrete blocks data from a building industry. These are given in $J = 24$ groups and $n = 50$ within-group blocks, where the groups were formed by combining all the modalities of: A = number of days passed from the production to the measurement (7 days, 28 days), B = production temperature (5°–10°C, 20°C); C = concrete typology (six in all).

In order to estimate the unknown, true strength ξ of a new block by considering its measure Y_0 obtained from the sclerometer, we have employed all the estimators described in Sects. 1–3. Then, we have compared them in terms of the RMSEP statistic with $K = 1$ and $N = 1{,}200$. The multilevel calibration estimator, with REML estimates of variance components, performs better than the others, having together with the within-group estimator, the lowest RMSEP value (Multilevel $= 2.032$, $df = 1{,}174$; Within-group $= 2.029$, $df = 1{,}151$; Classical $= 3.483$, $df = 1{,}197$; Oman's classical $= 4.954$, $df = 1{,}197$; Oman's contraction $= 4.954$, $df = 1{,}197$; Oman's ML $= 4.975$, $df = 1{,}197$).

5 Concluding Remarks

In the context of clustered measures, we have introduced the multilevel calibration estimator (MLVCE) for directly including the group membership information of an item requiring to be measured. Our approach substantially differs from the one Oman developed for repeated measurements. The MLVCE is based on cluster-specific parameters, which allow the available information about the cluster membership to be simply conveyed into the prediction process. Oman's estimators are based on population-averaged parameters, i.e. fixed effects, that should help provide the best possible prediction of a measure if nothing was known about it. Otherwise our estimator includes random–effects predictions in its variance, with respect to Oman's estimators. In future we should explore this very aspect.

A simulation study we carried out has highlighted that the MLVCE can perform very satisfactorily under specific sets of experimental conditions, in particular high levels of within-group correlation coefficient of the two measures. Nevertheless, several points would actually require more thorough examinations. The role of variance–covariance structure of random effects along with sample size should still be examined closely. Statistical properties of the MLVCE, for especially defining confidence limits, are still an open matter, which we are working on at present.

References

Brown, P. J. (1993). *Measurement, regression and calibration*. Oxford: Clarendon Press.
Cheng, C. L., & Van Ness, J. W. (1997). Robust calibration. *Technometrics 39*, 401–411.
Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Arnold Ed.

Oman, S. D. (1998). Calibration with random slopes. *Biometrika*, *85*, 439–449.
Osborne, C. (1991). Statistical calibration: A review. *International Statistical Review*, *59*(3), 309–336.
Pinheiro, C. J., & Bates, D. M. (2000). Mixed-Effects models in S and S-Plus. *Statistics and Computing*. New York: Springer-Verlag.
Pinheiro, C. J., Bates, D. M., DebRoy, S., Sarkar, D., & the R Core team. (2011). nlme: Linear and nonlinear mixed-effects models. R package version 3.1-101.
Salini, S. (2006). Robust multivariate calibration. In S. Zani, A. Cerioli, M. Riani, & M. Vichi (Eds.), *Data analysis, classification and the forward search* (pp. 217–224). Berlin: Springer.
Shiao, J., & Tu, D. (1995). *The jackknife and bootstrap*. New York: Springer.
Snjiders, T., & Bosker, R. (1999). *Multilevel analysis – An introduction to basic and advanced multilevel modelling*. London: Sage Publications.

Part V
Visualization of Relationships

Latent Ties Identification in Inter-Firms Social Networks

Patrizia Ameli, Federico Niccolini, and Francesco Palumbo

Abstract Social networks are usually analyzed through manifest variables. However there are social latent aspects that strongly qualify such networks. This paper aims to propose a statistical methodology to identify latent variables in inter-firm social networks. A multidimensional scaling technique is proposed to measure a latent variable as a combination of an appropriate set of two or more manifest relational aspects. This method, tested on an inter-firm social network in the Marche region (Italy), it is a new way to grasp social aspect with quantitative tools that could be implemented under several different conditions, using also other variables.

1 Introduction

In the last few years, the synergic integration between the organizational and statistical disciplines has been producing relevant scientific outputs. However this cross-disciplinary synergy can bring to new important evolutions, especially in the Social Network Analysis (SNA).

SNA studies the social resources exchange between actors and their relationships within a social system (Wasserman and Faust 1994). Ties measurement represents a relevant concern of SNA. Some authors postulate the existence of a latent 'social space' within which the presence of a tie between two actors is determined as function of some measures of [dis]similarity between the latent space positions of these actors. In this direction Leydesdorff et al. (2008) and Okada (2010) propose to use scaling models to determine the actors latent space positions and consequently the ties between actors.

In this framework, the paper aims to propose a statistical procedure to study ties in the latent space positions where actors are firms and ties are function of two or more measures of their relationships (Hoff et al. 2002).

As matter of fact, in the organizational science, when focusing on inter-firm social network many authors remark the complex nature of the social relationships. Grandori and Soda (1995) define the inter-firm network as 'a mode of regulating interdependence between firms, which is different from the aggregation of these

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_21,

units within a single firm and from coordination through market signals, and which is based on a co-operative game with partner-specific communication'. Particularly, social inter-firm networks can be defined as long term informal relationships among two or more organizations.

Nowadays in the emerging inter-organizational architectures, such as learning and visionary networks, relationships are built and planned more focusing on social basis instead than on formal agreements. In this context, partnerships' social latent aspects become the main ground for knowledge sharing and for long term performance.

Members of inter-firms social networks aim at sharing core values, knowledge, and, in some cases, they share also a vision. Topics like social capital, cohesiveness, and the embeddedness have a strong social component, that is not directly measurable but that can be identified through the observation of their manifest variables. Consequently the key research question concern the possibility to better study and quantify deep latent aspects of a social network, using manifest variables.

In social inter-firms networks typical examples of manifest measures are *trust*, *frequency of exchanges* among parties, and *reciprocity* (Jones et al. 1997). Many scholars indicate *trust* as fundamental ingredient for every typology of network or alliance and as facilitator of knowledge transmission (Nonaka 1991) and vision sharing. To have trust relationship is necessary to built it in a daily behavior (Lomi 1997, p. 214).

The *frequency* concerns how often specific parties exchange with one another (Jones et al. 1997, p. 917). Frequency is important for three reasons. First, frequency facilitates transferring tacit knowledge, second, frequent interactions establish the conditions for relational and structural embeddedness, which provide the foundation for social mechanisms to adapt, coordinate, and safeguard exchanges effectively and third, frequent interactions provide cost efficiency in using specialized governance structures (Williamson 1985, p. 60; Jones et al. 1997, p. 917).

Reciprocity 'transforms a unilateral supply relationship into a bilateral one' (Williamson 1985, p. 191; Jones et al. 1997, p. 922) and creates the perception of a similar 'destiny' with greater 'mutual interest' (Williamson 1985, p. 155; Jones et al. 1997, p. 922).

This paper proposes that the latent space can be defined as 'basic social relational embeddedness', and it is obtained as a function of the three chosen manifest variables: trust, frequency of exchanges among parties, and reciprocity. The basic social embeddedness is a component of the strength of the tie (Granovetter 1973), that is especially important for knowledge and vision sharing (Uzzi 1999).

The basic social embeddedness is identified in this work as the core part of the 'relational embeddedness': a component of the strength of the tie (Granovetter 1973), that is especially important for knowledge and vision sharing (Uzzi 1999). Relational embeddedness is an indicant of the motivational aspect of tie strength (Rindfleisch and Moorman 2001), essentially refers to the quality and depth of a single dyadic tie, it is fundamental mechanism of social governance, and captures the quality of dyadic exchanges and the behaviours exchange parties exhibit, such as trust, confiding, and information sharing (Jones et al. 1997, p. 924).

2 The Unique Social-Relational Variable

Scholars are familiar with the idea of *latent variable* in both organisational and in behavioral sciences. Latent variables refer to concepts that cannot be directly measured and are opposed to the observable variables. However, latent variables may be defined on the basis of properly identified sets of observable variables, called also manifest variables (Bartholomew 1987).

In the SNA framework, the network graphical visualization implies the identification of a *metric* space where actors and ties are represented. Generally the strength of the ties are represented in terms of distance between two actors or/and by the ties weight. In some cases suitable strength thresholds are defined: when strengths are lower than the threshold, the corresponding tie is omitted. However, the approach presented in the present paper assumes that ties are function of two or more measures and they only exist in latent space and not otherwise (Hoff et al. 2002).

The proposed procedure can be summarized in the following steps: (a) according to a set of two or more measures of relationships, actors are displayed in a latent metric space; (b) ties between actors depend on the distances between them in the latent (unobserved) space.

Formally we define the following data structure. Given a set of n statistical units and K manifest measures (variables), the notation δ_{ijk} indicates the general proximity measure between the statistical units i and j for the measure, which are arranged in K asymmetric $n \times n$ matrices $\{\Delta_1, \Delta_2, \ldots, \Delta_k, \ldots, \Delta_K\}$, where $(i, j) = 1, \ldots, n$. We remember that $\delta_{ijk} = \max$, $\forall k$ and if $i = j$ (diagonal elements), and that $\delta_{ijk} \geq 0$ and the value 0 indicates absence of any relationship, by definition.

The final aim of the paper is to identify and visualize ties in the latent space of the inter-firm social network; from a technical point of view, the problem consists in finding a good approximation of the 3-way proximity matrix by a $n \times n$ distance matrix.

2.1 Multidimensional Scaling for Ties Identification

This section shortly presents the MultiDimensional Scaling (MDS) basic principles and then it motivates the choice of the PROXSCAL model with respect to other scaling models. For sake of space, we do not go to deepen the PROXSCAL; interested readers can refer more specifically to Borg and Groenen (2005) for the MDS foundations and to Meulman and Heiser (2001) for the PROXSCAL model, more specifically.

The MDS aims at finding a configuration of n points into a p dimensional space. Generally, p is set equal to 2 in order to get graphic representation of the n points (Takane 2007). The scaling transformation of the proximity matrix has two advantages: (1) to summarize different proximity measures into one single distance; (2) to permit distances graphical representations into 2 and 3 dimensional spaces.

Moreover, it is worth noticing that the transformed relationship measures into metric measures can be displayed according to different approaches to visualize the network.

In the simplest two dimensional case, where $K = 1$, the MDS aims at defining a function $\varphi(\cdot)$ such that:

$$\varphi\left(\delta_{ij}\right) = d(x_i, x_j) + \epsilon_{ij} \tag{1}$$

In other words, $\varphi(\cdot)$ indicates a function that maps the dissimilarities δ_{ij} into a metric space, where the distance $d(\cdot)$ (generally the Euclidean distance) is defined. The quantity ϵ_{ij} indicates the residual.

Some alternative models have been proposed to deal with 3-Way data structures. In this paper we refer to the INDividual SCALing (INDSCAL) model, and more specifically to the PROXSCAL algorithm for 3Way dis(similarities) data structures. INDSCAL model is also referred to as weighted Euclidean model. Starting from K proximity matrices $\Delta_1, \Delta_2, \ldots, \Delta_k, \ldots, \Delta_K$, INDSCAL minimizes the Stress (squared Euclidean distances) defined by the following equation:

$$\sigma(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k, \ldots, \mathbf{X}_K) = \sum_k \sum_{i<j} \left(\delta_{ijk} - d_{ij}(\mathbf{G}\mathbf{W}_k)\right)^2, \tag{2}$$

where $\mathbf{G}$ indicates the common space $n \times n$ matrix and $\mathbf{W}_k$ represents the generic weighting matrix. The point coordinates are then defined as $\mathbf{X}_k = \mathbf{G}\mathbf{W}_k$ and the scaled distances as $d_{ij}(\mathbf{X}_k)$.

Differently from INDISCAL, PROXSCAL minimizes Normalized Stress σ_n:

$$\sigma_n(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k, \ldots, \mathbf{X}_K) = \frac{\sum_k \sum_{i<j} \left(\delta_{ijk} - d_{ij}(\mathbf{X}_k)\right)^2}{\sum_k \sum_{i<j} \delta_{ijk}^2}. \tag{3}$$

The main point in favor of PROXSCAL is that it works on the Euclidean distance and not on the squared distances. Avoiding the square transformation, it prevents putting more emphasis on large dissimilarities. The second, but not less important point, is the possibility to consider asymmetric (dis)similarity matrices. However, it is worth noticing that dissimilarity matrices are transformed in symmetric ones. Alternative scaling models could be taken into account, and this surely represents a research direction for future works. Last, the normalized stress is a relative measure that allows us to appreciate the overall quality of the solution and to make comparisons among several models.

3 Empirical Evidence

This section presents the output obtained on a small real dataset. The case study is an Italian inter-firms network in the Marche region (Ancona administrative district).

Nexus can be defined as an 'heritage network' as well as their firms are all located in a 'heritage area': a delimited territory (Vallesina) with a specific cultural identity. Network membership needs to share some values (mainly coming from the traditional farmers' culture of that area). The whole network consists of 25 firms, operating in different and potentially integrated field of activities (automotive and energy, consultancy, software house, clothing...), with a low level of reciprocal competition among actors. Firms to be officially included in Nexus need to share a collaborative mission, consequently relationships are mainly collaborative. Even if some official activities are periodically organized (such as meetings), the network is aimed to stimulate informal relationships and knowledge exchange among firms.

The latent space considered refers to the *basic social embeddedness*; the manifest measures are: *trust*, *frequency of exchanges among parties*, and *reciprocity*. The same questionnaire was given to a sample of eleven firms belonging to the network; using a scale $\{0, 1, \ldots, 5\}$, respondents have evaluated their partners on *trust*, *frequency of exchanges*, and *reciprocity*. For these three variables, each respondent was asked to indicate her/his feeling with respect to other actors in the net: scores indicate the measure of relationship measured on a non-linear similarity scale from 1 to 5. Higher values indicates high higher involvement degrees with other actors in the net. By definition we assumed that the self evaluation is equal to 5.

Data have been arranged into a three-way data matrix. Each slice has dimension 11×11, and the first row/column refer to Gruppo Loccioni, which has been the promoter of Nexus. Other firms are labelled with the capital letters 'B, C, ...'.

The matrix has been analyzed with the PROXSCAL procedure, and choosing a three dimensional solution. Generally researcher prefer the two-dimensional solution because it permits to get good graphical representations. However, the aim of the present paper is to visualize the scaled similarities using a network display.

In the following part of this section is summarized the PROXSCAL output of the SPSS package. To be consistent with the paper general aim, the *weighted Euclidean model* has been selected. This model defines the points co-ordinates in the common space as a weighted sum of the single space co-ordinates. Proximity transformations have been imposed on interval scale and applied across all sources simultaneously. The algorithm, starting from the Torgeson initial solution, reached the minimum value at $\sigma_n = 0.0277$, which represents a very good result. Table 1 displays the weighting matrix $\mathbf{W}$ having order 3×3 (three sources and three dimensions). It is worth noticing that all terms in $\mathbf{W}$ are positive. However, coefficients are very similar: in particular, *reciprocity* and *trust* have almost the same values. This coefficient signs concordance is consistent with the choice of an additivity model for defining the latent tie. It highlights that *reciprocity* and *trust* were perceived as equivalent concepts by the interviewed entrepreneur.

Figure 1 displays the statistical units in the 3D scaled space. All actors are connected with Gruppo Loccioni indicating the central role played by this firm inside Nexus.

Ties are function of the scaled distances among statistical units in the common space: $\sum_k d_{ij}(\mathbf{X}_k)$. The common space display (Fig. 1) allows to appreciate the reciprocal positions of the statistical units, and the network display (Fig. 2) shows

Table 1 Dimension weights

	Dimension Weights		
Source	Dim1	Dim2	Dim3
TRUST	0.397	0.334	0.347
REC	0.426	0.392	0.362
CONN	0.425	0.392	0.369

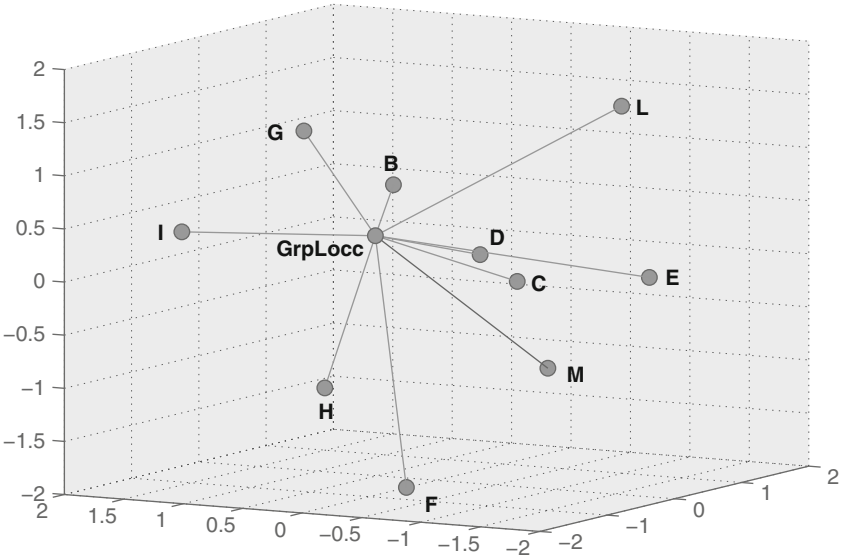


Fig. 1 Points in the 3D common space

the basic social embeddedness. Furthermore, vertices of the network represent the actors and the arcs thickness is function of the embeddedness.

Network latent ties display is represented in Fig. 2 it has been obtained using the Pajek© package. Scaled distances in the latent space among vertices (actors) represent the strength of the ties, it can be interpreted in terms of reciprocal 'social closeness'; at the same time the arcs thickness represents the basic social embeddedness as synthesis of the three considered variables. According to the most largely used techniques in SNA, when displaying the network proper thresholds can be defined to omit the visualisation of the weak ties.

Results show that the Nexus form is similar to a 'constellation of firms' where there is a focal enterprise, with a central role in the network. Regarding the basic social embeddedness, it is possible to observe that is deeper among the founders of the network, included the focal company. In fact, among the founders there is a high level of information exchange and knowledge creation. The firms that have been joined Nexus later, probably have some difficulties to integrate themselves with the founders. This means that the basic social embeddedness is stronger in the founders group than in the group of entrepreneurs that have been joined Nexus later.

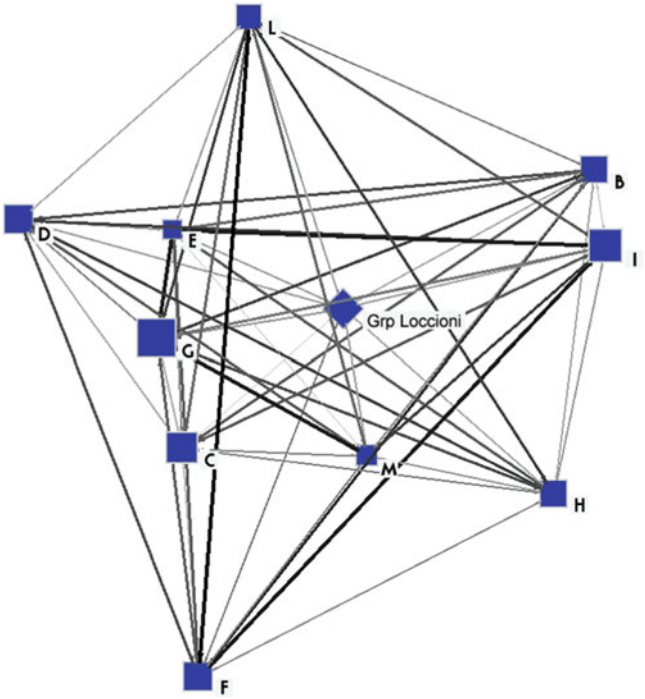


Fig. 2 Network display

This phenomenon has also a positive aspect as well as weak ties could be able to connect the network to others 'world' and to create new contacts with news firms (Granovetter 1983).

4 Conclusive Remarks: Possible Applications of the Proposed Method

Continuing to work at the network level, the proposed methodology can be used as a valid tool to study different kinds of social latent ties among organizations. *Trust*, *frequency*, *reciprocity* and *basic social embeddedness* are only an example of a set of manifest and latent variables: choosing sets of different manifest variables it will be possible to study other latent relational inter–organizational elements, such as network centrality and network density, that are important elements in a network analysis (Lomi 1997).

Measuring that kind of variables, it can help the management in the decisions concerning the investment in organizational structures that can facilitate the development of social relations. The method can also be viewed in an evolutionary perspective, comparing different measures of the latent tie over the time.

For this reason, this methodology is suitable to identify emerging significant networks' latent characteristics and consequently can offer new opportunities to study emerging inter-organizational settings that are becoming more relevant in the actual complex co-operative scenario, such as learning and visionary networks.

Researcher can also imagine applications of this method to identify latent ties, working at different organizational levels.

The research method can be used at the team level to analyze the main characteristics of the collaborative behavior between workers, such as the cohesiveness of the team, that are becoming more and more important especially in the perspective of the organizational learning studies (Senge 2006). At the organizational level, this method can be experimented to measure the latent relationship between the different units, providing organizational charters where the lines thickness connecting two units represents the measure of the latent tie studied.

Moreover, is it possible to imagine applications of the method at the macrosystemic level, especially for the authors that need to better understand the latent ties among organizations and their more relevant stakeholder groups.

References

Bartholomew, D. J. (1987). *Latent variable models and factors analysis*. New York, NY, USA: Oxford University Press, Inc.

Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling: Theory and applications* (2nd ed.). New York: Springer.

Grandori, A., & Soda, G. (1995). Inter-firm networks: Antecedents, mechanism and forms. *Organization Studies*, *16*(2), 183–214.

Granovetter, M. (1973). The strength of weak ties. *American Journal of the Sociology*, *78*(6), 1360–1380.

Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological Theory*, *1*, 201–233.

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *JASA*, *97*(460), 1090–1098.

Jones, C., Hesterly, W. S., & Borgatti, S. P. (1997). A general theory of network governance: Exchange conditions and social mechanism. *Academy oí Management Review*, *22*(4), 911–945.

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 la información*, *17*(6), 611–646.

Lomi, A., (1997). *Reti organizzative, teorie tecniche e applicazioni*. Bologna: Il Mulino.

Meulman, J. J., & Heiser, W. J. (2001). SPSS Inc., *SPSS Categories*©. Chicago: SPSS Inc.

Nonaka, I. (1991). The knowledge creating company. *Harvard Business Review*, *69*(6), 96–104.

Okada, A. (2010). Two-Dimensional centrality of asymmetric social network. In F. Palumbo, N. C. Lauro, & M. J. Greenacre (Eds.), *Data analysis and classification proceedings of the 6th conference of the ClaDAG* (pp. 93–100). Berlin Heidelberg: Springer.

Rindfleisch, A., & Moorman, C. (2001). The acquisition and utilization of information in new product alliances: A strength of ties perspective. *Journal of Marketing*, *65*, 1–18.

Senge, P. (2006). *La quinta disciplina*. Milano: Sperling & Kupfer Editori.

Takane, Y. (2007). Applications of multidimensional scaling in psychometrics. In C. R. Rao & S. Sinharay (Eds.). *Handbook of statistics, Vol. 26, Psychometrics*. USA: Elsevier B.V.

Uzzi, B. (1999). Embeddedness in the making of financial capital: How social relations and networks benefit firms seeking financing. *American Sociological Review*, *64*, 481–505.
Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, UK: Cambridge Univ. Press.
Williamson, O. E. (1985). *The economic institutions of capitalism – firms, markets, relational contracting*. New York: The Free Press.

A Universal Procedure for Biplot Calibration

Jan Graffelman

Abstract Biplots form a useful tool for the graphical exploration of multivariate data sets. A wide variety of biplots has been described for quantitative data sets, contingency tables, correlation matrices and matrices of regression coefficients. These are produced by principal component analysis (PCA), correspondence analysis (CA), canonical correlation analysis (CCO) and redundancy analysis (RDA). The information content of a biplot can be increased by adding scales with tick marks to the biplot arrows, a process called *calibration*. We describe a general procedure for obtaining scales that is based on finding an optimal calibration factor by generalized least squares. This procedure allows automatic calibration of axes in all forementioned biplots. Use of the optimal calibration factor produces graduations that are identical to Gower's predictive scales. A procedure for automatically shifting calibrated axes towards the margins of the plot is presented.

1 Introduction

Biplots are well-known graphical representations of multivariate data sets. Biplots have been described for most classical multivariate methods that are based on the singular value decomposition. While biplots do a good job at exploring multivariate data, they can still be improved in certain respects. In comparison with scatterplots, biplots are less informative, because the variables are represented by simple arrows without scales. The process of drawing a scale along a vector in a graph is called *calibration*. Probably the first example of a biplot with calibrated axes concerns the stork data analysed by Gabriel and Odoroff (1990). In this paper we describe a universal procedure for the calibration of axes in biplots. We first derive the basic calibration formulae for scatterplots and biplots in Sect. 2, the latter section summarizing the calibration formulae from Graffelman and van Eeuwijk (2005). Novel material is presented in Sect. 3 where we show the relationship with Gower's predictive scales, and in Sect. 4 where we treat axis shifting. Some examples (Sect. 5) and software references (Sect. 6) complete the chapter.

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_22,

2 Scatterplot and Biplot Calibration

The scatterplot is a well-known graphical representation of the relationship between a pair ($p = 2$) of quantitative variables. If there a more than two variables ($p > 2$), Gabriel's (1971) biplot obtained by principal component analysis is the natural generalization of the scatterplot. However, it is also possible to add extra (supplementary) variables to an existing scatterplot. A direction for a third variable that is optimal in the least squares sense can be found by the regression

$$\mathbf{b} = (\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'\mathbf{y}, \tag{1}$$

where $\mathbf{y}$ is a third quantitative variable to be "added" to an existing scatterplot, and $\mathbf{F}$ an $n \times 2$ matrix containing the two variables already represented in the scatterplot. All variables are assumed to be centered. We obtain a 2×1 vector of regression coefficients. This defines a direction from $(0, 0)$ to (b_1, b_2) in the scatterplot that is "best" to represent $\mathbf{y}$. A practical problem arises: how can we draw a scale in the natural units of $\mathbf{y}$ along this direction? We call this a *calibration* problem. Stated more generally, the question is how any direction ($\mathbf{g}$) in the scatterplot can be calibrated with an optimal scale for $\mathbf{y}$. Let scalar α be the *calibration factor* that determines the graduation along vector $\mathbf{g}$. The vector $\tilde{\mathbf{y}}$ represent the data values recovered when the data points are projected perpendicularly onto vector $\mathbf{g}$, and is given by $\tilde{\mathbf{y}} = (1/(\alpha\mathbf{g}'\mathbf{g}))\mathbf{F}\mathbf{g}$. We minimize the sum of squared errors ($\mathbf{e} = \mathbf{y} - \tilde{\mathbf{y}}$) with respect to α. Graffelman and van Eeuwijk (2005) obtained the optimal scaling factor:

$$\alpha = \frac{\mathbf{g}'\mathbf{F}'\mathbf{F}\mathbf{g}}{\mathbf{y}'\mathbf{F}\mathbf{g}\mathbf{g}'\mathbf{g}}. \tag{2}$$

If the regression coefficients in (1) are used as the direction to represent the third variable, then (2) simplifies to $1/\mathbf{g}'\mathbf{g}$. In most practical applications the direction supplied by the regression coefficients will be the preferred choice. However, sometimes other directions might be of interest, e.g., when a second vertical axis for a third variable is desired in the right margin of a scatterplot. One usually wishes to represent nice and equally spaced values $(0, 10, 20, \ldots)$ along the axis, the particular values depending on the range of variation of the variable in the sample. With $\alpha\mathbf{g}$ representing a unit increment, the $(m \times 2)$ matrix of tick mark positions for a set of values $\mathbf{t} = (t_1, \ldots, t_m)$ is obtained by:

$$\mathbf{M} = \alpha\mathbf{t}\mathbf{g}', \tag{3}$$

where the values in $\mathbf{t}$ must be centered and/or standardized in the same way as $\mathbf{F}$ was initially centered and/or standardized. If weights are used in the analysis of the data, then it is natural to compute the calibration by using weighted or generalized least squares (GLS). If GLS is used, and we minimize $\mathbf{e}'\mathbf{A}\mathbf{e}$ with $\mathbf{A}$ a symmetric positive definite $n \times n$ matrix of weights, then the optimal calibration factor is found to be:

$$\alpha = \frac{\mathbf{g}'\mathbf{F}'\mathbf{A}\mathbf{F}\mathbf{g}}{\mathbf{y}'\mathbf{A}\mathbf{F}\mathbf{g}\mathbf{g}'\mathbf{g}}. \tag{4}$$

Equations (3) and (4) form the basic calibration results that are easily adapted for their use in biplots. In general, biplots are factorizations of a data matrix $\mathbf{X}$ as $\mathbf{X} = \mathbf{F}\mathbf{G}'$, where data matrix $\mathbf{X}$ is typically approximated by using the first two columns of $\mathbf{F}$ and $\mathbf{G}$ only. The scatterplot is just a particular case of this factorization with $\mathbf{X} = \mathbf{F}\mathbf{I}'$. The optimal factorization of $\mathbf{X}$ is obtained by the singular value decomposition. A biplot is in fact a scatterplot of the row markers in $\mathbf{F}$, and the variables are represented by plotting the rows of $\mathbf{G}$ as arrows in the same plot. Calibration of a biplot axis is possible by applying (4) thereby substituting for $\mathbf{F}$ the $n \times 2$ matrix of biplot coordinates (the row markers), for $\mathbf{g}$ the particular biplot axis (a row of $\mathbf{G}$) we wish to calibrate and for $\mathbf{y}$ the original data vector of the variable represented by biplot axis $\mathbf{g}$. The main difference between calibrating a scatterplot and a biplot resides in the proper definitions of weight matrix $\mathbf{A}$ and vector $\mathbf{t}$. Matrix $\mathbf{A}$ depends on the particular multivariate method used to construct the biplot, and also on the matrix we wish to represent. E.g., in PCA we will have $\mathbf{A} = \mathbf{I}$, in CA $\mathbf{A} = \mathbf{D}_c$ or $\mathbf{D}_r$, in CCO $\mathbf{A} = {\mathbf{R}_{xx}}^{-1}$ or ${\mathbf{R}_{yy}}^{-1}$ (if we wish to calibrate vectors with a correlation scale ranging from -1 to 1) and in RDA $\mathbf{A} = \mathbf{I}$ or $\mathbf{R}_{xx}$.

3 Relationship with Gower's Predictive Scale

The topic of calibration of biplot axes has been addressed by Gower and Hand (1996). With a different rationale, the latter authors obtain the tick mark positions (Gower and Hand 1996 [p. 16]; Gardner-Lubbe et al. 2009 [p. 31]):

$$\frac{\tilde{\mu}\mathbf{e}'\mathbf{V}}{\mathbf{e}'\mathbf{V}\mathbf{V}'\mathbf{e}}, \tag{5}$$

with $\tilde{\mu} = \mathbf{t}$ and where $\mathbf{e}'\mathbf{V}$ corresponds to a row in a matrix of eigenvectors, which corresponds to $\mathbf{g}'$ in the notation of this paper. We thus have

$$\frac{\tilde{\mu}\mathbf{e}'\mathbf{V}}{\mathbf{e}'\mathbf{V}\mathbf{V}'\mathbf{e}} = \frac{1}{\mathbf{g}'\mathbf{g}}\mathbf{t}\mathbf{g}' = \alpha\mathbf{t}\mathbf{g}' = \mathbf{M}. \tag{6}$$

This shows the equivalence between Gower's predictive scales and the scales obtained by least squares in this paper.

4 Shifting of a Biplot Axis

When many vectors in a biplot are calibrated, the plot tends to be cluttered up by many scales passing through the origin, and it becomes difficult to make any sense out of it. It is a good idea to shift calibrated axes towards the margins. This way,

all information on the measurement scales is moved towards the margins of the plot, while leaving the cloud of points in the biplot unaffected. In fact, this is also what we do in an ordinary scatterplot: units, labels and tick marks are represented in the (left and bottom) margins of the plot. The first biplots with such shifted axes seem to stem from Graffelman and van Eeuwijk (2005). Bläsius et al. (2009) give some suggestions to shift axes in an automated manner. Here we develop automatic shifting of biplot axes in more detail, using matrix notation. The shift of a biplot axis can be described by vector **d** that is orthogonal to biplot vector **g**. Vector **d** can be set by hand, or some sensible choices can be calculated automatically. In particular, **d** could be chosen in such a way that all data points fall whether below or above the shifted axis. This requires determining the maximum and the minimum of the data points in the direction orthogonal to **g**. In order to find the extremes, we compute possible shift vectors $\mathbf{d}_i$ for all data points:

$$\mathbf{d}_i = \mathbf{f}_i - \frac{\mathbf{g}'\mathbf{f}_i}{\mathbf{g}'\mathbf{g}}\mathbf{g}, \tag{7}$$

where $\mathbf{f}_i$ is a row of matrix **F**. Next, we determine which $\mathbf{d}_i$ has the largest norm for all points "above" **g**, and which $\mathbf{d}_i$ has the largest norm for all points "below" **g**. Whether a point is "below" or "above" can be determined by an auxiliary clockwise

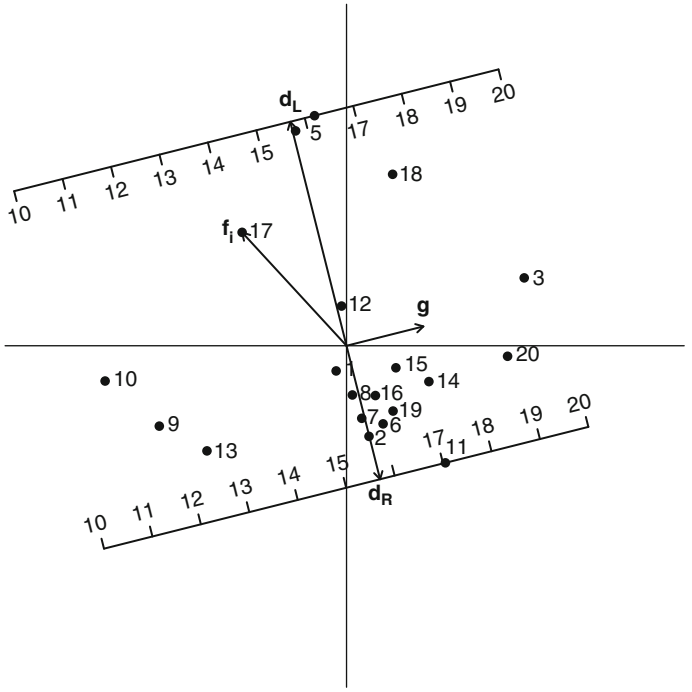


Fig. 1 Shifting biplot axis **g** towards the most extreme data points with shift vectors $\mathbf{d}_L$ and $\mathbf{d}_R$

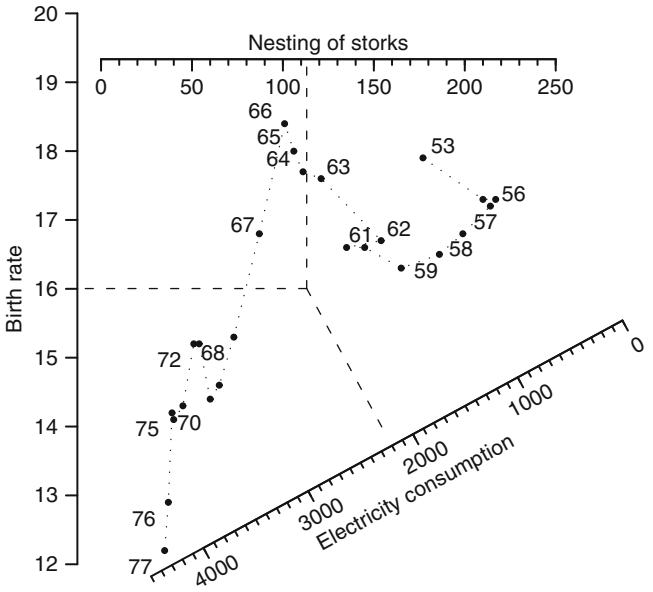


Fig. 2 Three-variable scatterplot of the stork data. The *dashed line* indicates the means of the variables

rotation of the data points over an angle as large as the angle between $\mathbf{g}$ and (1,0), and looking at the sign of the rotated coordinates. We call the maximal norm vectors $\mathbf{d}_R$ and $\mathbf{d}_L$, with suffix R and L for "right" and "left" in the direction of $\mathbf{g}$ (See Fig. 1). The final tick mark positions for the two shifted axes then become

$$\mathbf{M}_L = \alpha \mathbf{t}\mathbf{g}' + \mathbf{1}\mathbf{d}_L' \quad \text{and} \quad \mathbf{M}_R = \alpha \mathbf{t}\mathbf{g}' + \mathbf{1}\mathbf{d}_R'. \tag{8}$$

Depending on the particular data set, the most convenient position for the shifted axis can be chosen. With the current definition of the shift vectors $\mathbf{d}_L$ and $\mathbf{d}_R$, the most extreme data point perpendicular to $\mathbf{g}$ is exactly on the calibrated axis (e.g., observations 5 and 11 in Fig. 1). It is sensible to stretch these vectors slightly, so that all points are really above (or below) the corresponding axis. This shift procedure is implemented in version 1.6 of the R-package `calibrate` (Graffelman 2009).

5 Examples

In this section we look at a few examples of calibrated scatterplots and biplots, using the stork data from Gabriel and Odoroff (1971). This is a Danish data set where the frequency of nesting storks, birth rate and electricity consumption were registered for a period of 25 years (1953–1977). The forementioned paper contains

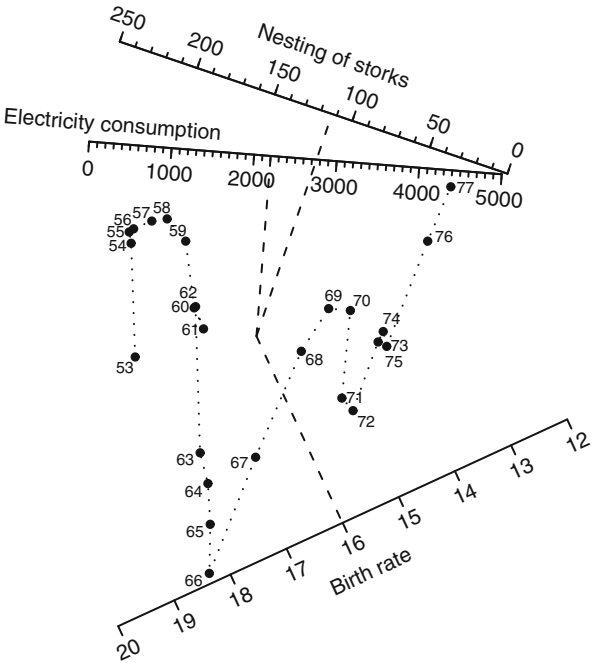


Fig. 3 PCA Biplot of the stork data, based on the correlation matrix. The *dashed line* indicates the means of the variables. Goodness of fit: Birth rate 0.997, electricity consumption 0.979, nesting frequency 0.990

a calibrated biplot of this data set. Figure 2 shows an alternative graphical representation of the same data, a scatterplot of birth rate versus frequency of nesting, where electricity consumption has been added as a supplementary variable by regression. The goodness of fit of the representation is high, as 94.63% of the variance of electricity consumption can be explained by the regression onto stork nesting and birth rate. The over all goodness of fit of the representation is therefore $100 \cdot (1.0 + 1.0 + 0.9463)/3 = 98.21\%$. This is almost as good as a biplot of the data obtained by a correlation-based PCA (98.88%). Figure 2 is a nearly perfect representation of how the three variables change over time, showing an overall decrease in birth rate and frequency of nesting, and an increase in electricity consumption. Shorter periods with higher and lower rates of change in the variables are easily identified. The figure is a non-standard representation of the data, which is not trivial to obtain. Depending on the shape of the cloud of points, a good position for each axis and good scale values must be chosen such as to make the graph as informative as possible.

A drawback of Fig. 2 is that the correlation structure of the variables is not very well represented, partly because of the forced orthogonality between birth rate and frequency of nesting. This aspect is improved if we make a calibrated PCA-biplot of the data, which is shown in Fig. 3. The sharp angles between the axes indicate that

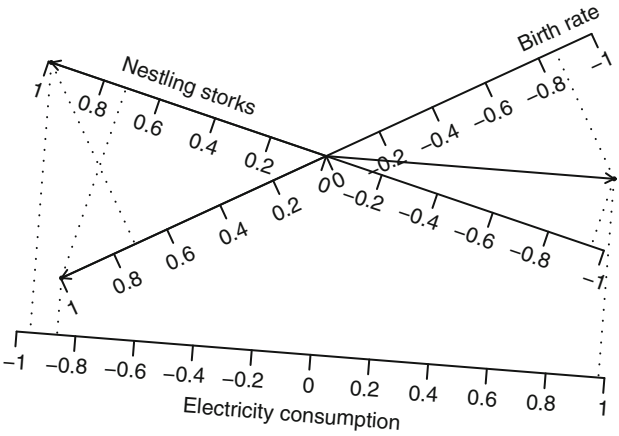


Fig. 4 PCA Biplot of the correlation matrix with a shifted axis for electricity consumption. Goodness of fit over 0.999 for all variables

all variables are correlated. Because the scale for electricity consumption runs in the opposite direction of those of the other two variables, correlations with electricity consumption are negative. For recovery of the original data matrix, Figs. 2 and 3 are equally well suited. For inferring the correlation structure, Fig. 3 is to be preferred.

However, the approximation of the correlations by angles is not optimized in PCA. PCA approximates the correlation matrix by scalar products, and the latter approximation is optimal in the least squares sense. It has the drawback that we also approximate the diagonal of ones. Figure 4 is a calibrated biplot of the correlation matrix of the three variables, and shows the approximation of the correlations by scalar products. The biplot arrow for electricity consumption has been shifted in order not to mess up the display. All correlations can now be read off by projection. The overall goodness of fit of the representation of the correlation matrix is 99.98%.

6 Software

R-package `calibrate` has been developed by the author for the calibration of axes in biplots and scatterplots. The package can be downloaded from the R project's home page (www.r-project.org). The package is extensively documented with examples of calibrated biplots for PCA, CA, CCO and RDA. Package `calibrate` provides a workhorse routine that computes calibration results and draws calibrated axes one by one. Its force resides in its generality: it can calibrate any scatterplot or biplot axis, and it can shift these. The idea behind the package is to create a calibrated axis by using an R-script that contains detailed instructions for tailoring the calibration of each axis.

`BiplotGUI` (Gardner-Lubbe et al. 2009) is alternative package for biplot construction and calibration with a graphical user interface and facilities for

representations in three dimensions. It also covers non-linear calibrations but seems limited to particular multivariate methods (PCA, CVA, MDS, Procrustes analysis). This package creates calibrated biplots by manipulation on-screen with a mouse or other pointing device.

References

Blasius, J., Eilers, P. H. C., & Gower, J. (2009). Better biplots. *Computational Statistics & Data Analysis*, *53*(8), 3145–3158.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, *58*(3), 453.

Gabriel, K. R., & Odoroff, C. L. (1990). Biplots in biomedical research. *Statistics in Medicine*, *9*(5), 469–485.

Gardner-Lubbe, S., le Roux, N., & la Grange, A. (2009). `BiplotGUI`: Interactive biplots in `R`. *Journal of Statistical Software*, *30*(12), 1–37.

Gower, J. C., & Hand, D. J. (1996). *Biplots*. London: Chapman and Hall.

Graffelman, J. (2009). `calibrate`: Calibration of biplot and scatterplot axis. `R` package version 1.6. http://cran.R-project.org/package=calibrate.

Graffelman, J., & van Eeuwijk, F. (2005). Calibration of multivariate scatter plots for exploratory analysis of relations within and between sets of variables in genomic research. *Biometrical Journal*, *47*(6), 863–879.

Analysis of Skew-Symmetry in Proximity Data

Giuseppe Bove

Abstract Skew-symmetric data matrices can be represented in graphical displays in different ways. Some simple procedures that can be easily performed by standard software will be proposed and applied in this communication. Other methods based on spatial models that need ad hoc computational programs will be also reviewed emphasizing advantages and disadvantages in their applications.

1 Introduction

Skew-symmetric data matrices occur in different fields of application as: debits/credits balance data, preferences, results of matches in a tournament, etc. Frequently they are derived as the skew-symmetric component of an asymmetric data matrix representing flow data, import/export data, confusion rates, etc. Even in this second situation, in this paper we assume that the main interest is to analyse separately the skew-symmetric component from the symmetric component.

Graphical displays can help to detect asymmetric relationships by associating geometrical entities to the information contained in the data matrix. Some simple procedures will be proposed in the next section and applied to a small data set concerning musical preferences. An advantage of the proposals is their easy application even with standard no statistical software available in every personal computer. In the third section it is shown how it is possible to analyse the size and the sign of skew-symmetry by standard statistical software containing symmetric multidimensional scaling routines. Finally, in the last section some other models for multidimensional graphical representation that need ad hoc computational programs are also briefly reviewed.

2 Simple Descriptions of Skew-Symmetry

A small asymmetric data matrix concerning choice probabilities among four music composers (labelled B, H, M, S) was reported in Takane (2005, Table 1), where each entry is the proportion of times (p_{ij}) row composer i is preferred

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_23,

Table 1 Log transformations of choice probabilities ratios among four music composers

Composers	B	H	M	S
B	0	2.143	.974	2.143
H	−2.143	0	−1.758	−.189
M	−.974	1.758	0	1.457
S	−2.143	.189	−1.457	0

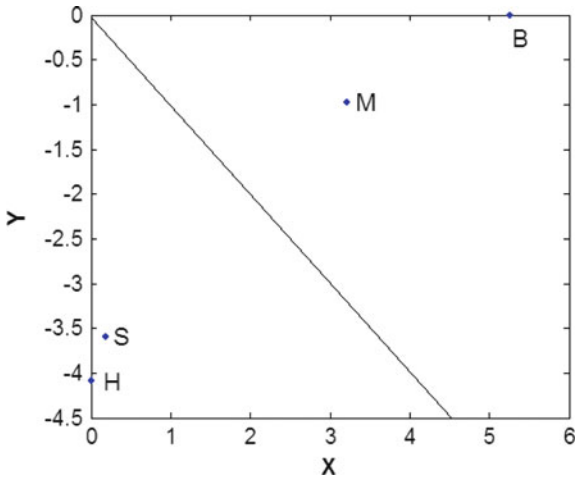


Fig. 1 Composer positive and negative marginal skew-symmetry

to column composer j. Entries in the skew-symmetric matrix presented in the following Table 1 (also reported in Takane 2005) are obtained by the transformation $n_{ij} = \log(p_{ij}/p_{ji})$ of the observed proportions. When n_{ij} is positive the row composer is more frequently preferred to the column composer and viceversa when n_{ij} is negative.

Skew-symmetric preference tables are usually studied by specific models like Thurstone or Bradley-Terry-Luce models, but in this paper we take a general explorative approach to skew-symmetry focusing on graphical capabilities of the methods.

In a similar manner to what is usually done for asymmetric proximity matrices, we could proceed to represent row and column totals of the data matrix. However, for skew-symmetric matrices row and column totals corresponding to the same object i sum to zero, so it seems preferable to analyse the prevalence of positive or negative skew-symmetry for each row of the matrix. A simple method can be based on a diagram where each composer is represented by a point with coordinates (x, y) that are sums, respectively, of the positive and negative entries in the row associated to that composer. The diagram obtained for Table 1 is depicted in Fig. 1.

In the diagram, points above the line $y = -x$ represent composers with a prevalence of positive skew-symmetries (*preferred*), points under the line represent composers with a prevalence of negative skew-symmetries (*not preferred*). In the

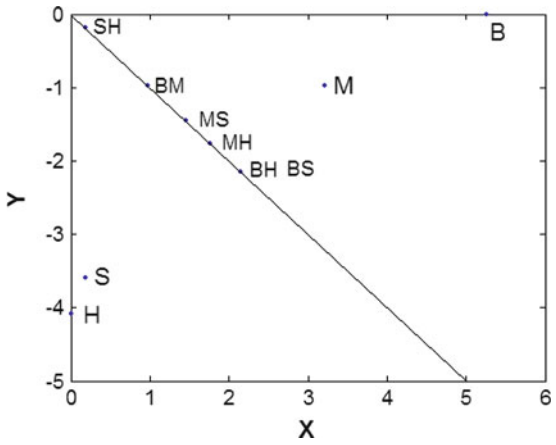


Fig. 2 Skew-symmetries of pairs of composers

same diagram we can also represent the 6 pairs of composers (i, j) with coordinates given by $x = |n_{ij}|$ and $y = -|n_{ij}|$, where $|\bullet|$ is the absolute value. If l_i and l_j are labels for i, j we adopt the convention to represent the pair with $l_i l_j$ if $n_{ij} \geq n_{ji}$ and $l_j l_i$ otherwise. The result is depicted in Fig. 2.

As it was expected composer B is more frequently preferred to the others, composer M is more frequently preferred to composers H and S. Largest skew-symmetries regard pairs BH and BS. Pairs with small skew-symmetries are depicted close to the origin.

The number of pairs to be represented increases with the number n of rows of the skew-symmetric matrix (it is $n \times (n-1)/2$), hence for large n only selected subgroups of pairs should be represented in the diagram.

3 Graphical Display of Size and Sign of Skew-Symmetry

Sizes of skew-symmetries $|n_{ij}|$ can be represented by distances in a *low-dimensional* Euclidean space (usually two-dimensional) between only n points by the model:

$$f(|n_{ij}|) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)} + e_{ij} \tag{1}$$

where f is a chosen data transformation function, $\mathbf{x}_i, \mathbf{x}_j$ are vectors of coordinates for the rows i and j and e_{ij} is an error term. The method can be easily applied by standard statistical software containing symmetric multidimensional scaling routines, and it allows to incorporate non metric approaches and external information regarding rows (Bove 2006). Figure 3 represents the result obtained by performing a symmetric MDS of composers data (*Stress-1* $= 0.03$). Large distances represent skew-symmetries between pairs of composer (B,H), (B,S), (M,H) and (M,S), small distances represent skew-symmetries between pairs (H,S) and (B,M).

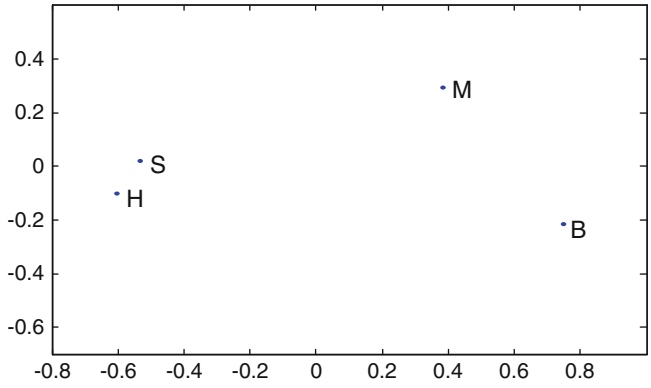


Fig. 3 Size of skew-symmetries of pairs of composers

The sign of skew-symmetry cannot be analysed in the diagrams obtained by model (1). A simple way to fill the gap can be based on the idea to draw circles around points of Fig. 3, originally proposed by Okada and Imaizumi (1987) to represent skew-symmetry (jointly with symmetry).

In our example, circles around points can represent the sign of skew-symmetry when compared in ordered pairs of points. That is, when circle around point i is larger than circle around point j than estimate of skew-symmetry n_{ij} is positive and estimate of skew-symmetry n_{ji} is negative. We propose to estimate the radii r_i of the circles by the following model

$$d_{ij} = (r_i - r_j) + e_{ij}, \tag{2}$$

where $d_{ij} = 1$ if n_{ij} is positive and $d_{ij} = -1$ if n_{ij} is negative. A least squares solution for the r_i's in (2), provided by Mosteller (1951) in the context of Thurstone case V scaling, is

$$\hat{r}_i = \frac{1}{n} \sum_{j=1}^{n} d_{ij}$$

with $\sum_{i=1}^{n} \hat{r}_i = 0$, being matrix $D = [d_{ij}]$ skew-symmetric. Any translations $\hat{r}_i + c$ by a constant c is equivalent to the initial solution, but it is convenient to choose only between solutions with non negative $\hat{r}_i'$s because they represent radii. In this application we choose the unique solution having $min(\hat{r}_i) = 0$.

Figure 4 represents the result obtained adding in Fig. 3 the radii obtained for composers data with model (2). The radii were rescaled so that the circles could be represented most conveniently in the configuration. Composer B has the largest radius that means he has always positive skew-symmetry with all the others composers. Composer H has a null radius that means he has always negative skew-symmetry with all the others composers.

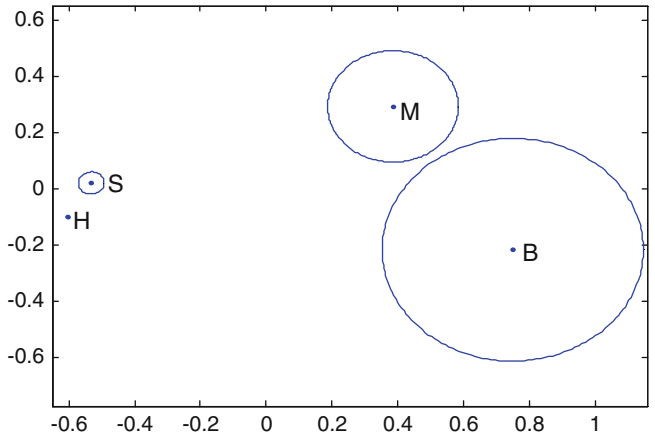


Fig. 4 Size of skew-symmetries of pairs of composers and circles

In the following section we will consider other important multidimensional models that allow to avoid the inconvenience of model (1). These models need 'ad hoc' computational programs.

4 Multidimensional Representation of Skew-Symmetry

The first multidimensional method for representing skew-symmetry was proposed in a pioneering paper by Gower (1977). It is based on a scalar product like model of the following type

$$n_{ij} = \mathbf{x}_i' \Delta \mathbf{J} \mathbf{x}_j + e_{ij} \tag{3}$$

where the vectors of coordinates $\mathbf{x}_i, \mathbf{x}_j$, the matrix $\Delta = diag(\delta_1, \delta_1, \delta_2, \delta_2, \ldots)$ and the block diagonal matrix $\mathbf{J}$, 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), are obtained by the singular value decomposition (SVD) of the skew-symmetric data matrix (see also Bove 2010); and e_{ij} is an error term. For this method 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 counter-clockwise).

Figure 5 represents the Gower diagram obtained for composers data. Now the size of skew-symmetry regarding, for example, the pair of composers (B,M) is proportional to the area of the depicted triangle, while the sign is positive counter-clockwise, that is composer B is more frequently preferred to composer M. At the same manner, we can see that composer B is more frequently preferred to

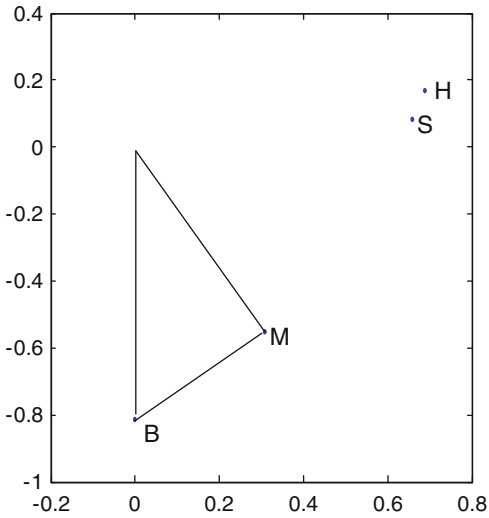


Fig. 5 Gower diagram of skew-symmetries of pairs of composers

all the other composers, and triangle areas are especially large with composers H and S. Since the four points are almost aligned, one-dimensional models (like e.g. Bradley-Terry-Luce model) for preference data could fit quite well in this case.

Rotational indeterminacy characterizing Gower method can allow to isolate different systems of asymmetric relationships in different planes when we fit more than two dimensions by opportune rotation methods (Rocci and Bove 2002).

The analysis of distances in a diagram is easier than the analysis of areas of triangles, so Borg and Groenen (2005) proposed to represent skew-symmetry by the model:

$$n_{ij} = sign(\mathbf{x}_i' \mathbf{J}' \mathbf{x}_j) \sqrt{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)} + e_{ij} \tag{4}$$

where vectors $\mathbf{x}_i, \mathbf{x}_j$, matrix $\mathbf{J}$ and the term e_{ij} are defined as in (3). Distances between points estimate the size of skew-symmetry and the direction of rotation provide 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}. Borg and Groenen (2005, pp. 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, we notice that Borg and Groenen say their algorithm can be quite sensitive to local optima.

Figure 6 shows what we could obtain by applying model (4) to composers data. The half-plane that point B determines with the origin in clockwise direction is depicted in the diagram. All points that represent the other composers fall into this half-plane, this means that composer B has positive skew-symmetries with all the others.

The proposal of Borg and Groenen allows to avoid the inconveniences of distance models applied only to the size of skew-symmetry.

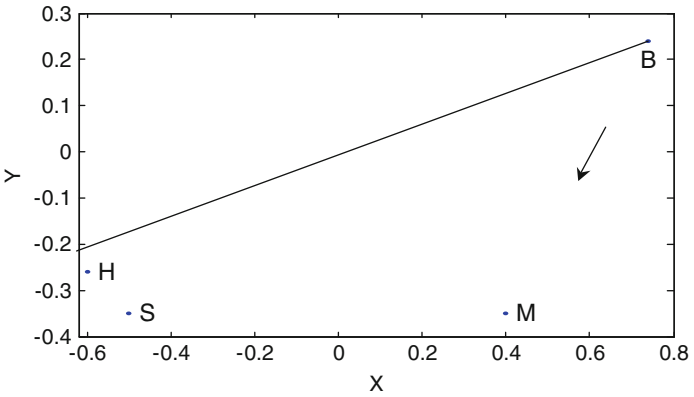


Fig. 6 Example of Borg and Groenen diagram of skew-symmetries of pairs of composers

A review of these and other multidimensional methods for skew-symmetry, including the three-way case, is provided in Bove (2010). Most methods can be applied only by ad hoc computational programs.

5 Conclusions

Taking the point of view of a standard user we have proposed some simple procedures to represent in a diagram skew-symmetric data matrices by largely available software.

When statistical software including MDS routines is also available, representations of size of skew-symmetry are easily obtainable and we have shown that the sign of skew-symmetry can also be included by drawing circles around points.

Some multidimensional methods based on spatial models that need ad hoc *computational programs* were also reviewed emphasizing advantages and disadvantages in their applications. Future developments could concern the study of distance like-models and their application to the three-way case.

References

Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling. Theory and applications*. Second edition. New York: Springer.

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 Verlag.

Bove, G. (2010). Methods for the analysis of Skew-symmetry in asymmetric multidimensional scaling. In H. Locarek-Junge & C. Weihs (Eds.), *Classification as a tool for research. Studies in classification, data analysis, and knowledge organization* (pp. 271–278). Berlin: Springer Verlag.

Gower, J. C. (1977). The analysis of asymmetry and orthogonality. In J. R. Barra, et al. (Ed.), *Recent developments in statistics* (pp. 109–123). Amsterdam: North Holland.

Mosteller, F. (1951). Remarks on the method of pair comparisons: I. the least squares solution assuming equal standard deviation and equal correlations. *Psychometrika*, 16, 3–9.

Okada, A., & Imaizumi, T. (1987). Non metric multidimensional scaling of asymmetric similarities. *Behaviormetrika*, 21, 81–96.

Rocci, R., & Bove, G. (2002). Rotation techniques in asymmetric multidimensional scaling. *Journal of Computational and Graphical Statistics*, 11, 405–419.

Takane, Y. (2005). Scaling asymmetric tables. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics for behavioral sciences* (pp. 1787–1790). Chichester: Wiley.

Social Stratification and Consumption Patterns: Cultural Practices and Lifestyles in Japan

Miki Nakai

Abstract The aim of this paper is to examine the relationship between cultural consumption and social stratification. Based upon a nationally representative 2005 Japanese sample (N = 2,915), we uncovered the association between a wide range of cultural capital and social class in Japan. In doing so, we re-examined conventional occupational schemes and developed a detailed occupational classification. Correspondence analysis revealed that both men and women who are well-educated and have a higher occupational position have more cultural capital. The results also indicate gender-specific cultural consumption patterns. For women, highbrow culture is important for distinguishing themselves and maintaining social position. In contrast, highbrow culture is defined as an irrelevant waste of time for men of higher position and instead business culture, which is characterized by a mixture of enterprise and rationality, prevails.

1 Introduction: Consumption Patterns and Social Stratification

In sociological theory focusing on the relationship between cultural consumption and social stratification, it has been maintained that lifestyle is an expression of one's class position. Lifestyle of the higher status group has been characterized by cultural refinement, i.e., highbrow taste is a feature of the well-educated elite in the society (Bourdieu 1979, Veblen 1899). Other sociologists emphasize that, in modern societies, the cultural taste of the higher status group is inclusive and broad, not snobbish or exclusive such that other non-elite classes are excluded. Empirical research shows that those of high status not only participate more regularly in high-status activities but also tend to participate more often in a lot kinds of leisure activities (Peterson and Simkus 1992, DiMaggio 1987). This is called the cultural omnivore-univore hypothesis. In either hypothesis, it is assumed that differential participation in cultural practices between individuals is related to their social status. A substantial number of empirical studies have been found evidence supporting differences in cultural activity between the higher educated and those of lower educational background (e.g., Kataoka 1992). We also have so far found that the cultural

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_24,

preferences differ among people from different social groups (Nakai 2005, 2008). These differences in cultural consumption patterns can be attributed to educational and family background.

It is also claimed that being health-conscious is also associated with social class. People from different social groups have different levels of concern over chemicals and products that may be unhealthy in their food for example, which in turn is associated with health inequality.

However, some sociologists claim that in postmodern consumer societies, consumption patterns no longer act to structure social class. Moreover, in Japan, the association between social stratification position and cultural taste remains controversial despite numerous studies. Many scholars have questioned the validity of the existence of the link between highbrow culture and people of higher status, especially among males, since in Japan quite a few educated high status men enjoy popular culture activities, such as karaoke and reading comic books, regardless of their education and social status.

The controversy over the relationship between social stratification and cultural consumption may be attributed in part to the class schemes which are conventionally used for analysis. Conventional class categories allegedly poorly correlated with actual life conditions because the bit conventional classifications group occupations together that differ substantially in potential life chances (Weeden and Grusky 2005).

Another reason for this controversy is the conceptualization of cultural taste, i.e., highbrow culture versus popular- or lowbrow culture. Many empirical studies have traditionally operationally defined cultural taste by using only limited consumption fields that are presumed to be significant for the reproduction of the stratification system.

The aim of this paper is to make social distinctions and inequalities between social groups visible by classifying cultural practices and lifestyles as well as position in the class structure. In doing so, we take wider cultural consumption into account. As to class scheme, we devised detailed occupational categories.

2 Methods

2.1 Data

The data is from a nationally representative 2005 survey of Social Stratification and Social Mobility conducted in Japan. Of approximately 13,000 men and women sampled, aged 20–69, 5,742 (2,660 men and 3,082 women) were successfully interviewed.[1] This data provides a wide range of socio-economic characteristics, information about values, family life, and leisure time activities. We analyzed 2,915

[1] The response rate was not very high and younger males may be slightly under-represented.

(1,317 men and 1,598 women) respondents, who were asked detailed questions concerning a wide range of cultural consumption practices to portray the structure of cultural participation and lifestyles.[2]

2.2 Measurement of Cultural Consumption and Lifestyle

We focused on three chief aspects (or domains) of cultural lifestyle: cultural practices, leisure activities and consumption patterns. The following 17 variables in cultural consumption of these three domains have been selected for investigation:

- Cultural practices; (1) classical music performances and concerts, (2) museums and art exhibitions, (3) karaoke, (4) playing some sports, (5) public library use, (6) reading sports journals or women's weeklies, (7) reading novels, or books about history;
- Leisure activities; (8) taking culture lessons, (9) language study, (10) travel abroad, (11) volunteer activities;
- Consumption patterns; (12) credit card purchasing, (13) online shopping and booking, (14) mail order catalog shopping, (15) eating out, drawing upon Restaurant Guides, (16) purchase domestically-produced groceries, and (17) purchase organic food.

2.3 Occupational Scheme

As to our occupational stratification scheme, we subdivided the conventional class grouping into 23 categories in terms of educational credentials, which reflect skill levels, size of organizations, worker functions, and sectoral differences (representing private sector or public sector).[3] We consequently devised the following disaggregated occupational categories: (1) professionals with higher educational **(prof h c)**, (2) cultural and artistic producers **(cul pro)**, (3) technicians **(techn)**, (4) teachers **(teach)**, (5) healthcare professionals **(hlth prof)**, (6) professionals not elsewhere classified **(prof(n.e.c.))**, (7) managers in large firm **(mgrl l)**, (8) managers in small firm **(mgrl s)**, (9) government officials **(gvrn offc)**, (10) clerical in large firm **(cler l)**, (11) clerical in small firm **(cler s)**, (12) sales in large firm **(sale l)**, (13) sales in small firm **(sale s)**, (14) proprietors **(proprt)**, (15) skilled manual in large firm **(skil l)**, (16) skilled manual in small firm **(skil s)**, (17) self-employed skilled manual **(s-e skil)**, (18) semi-skilled manual in large firm **(sskl l)**, (19) semi-skilled manual in small firm **(sskl s)**, (20) unskilled work in large firm **(uskl l)**, (21) unskilled work in small firm **(uskl s)**, (22) independent unskilled work **(ind uskl)**, and (23) farmers **(farm)**.

[2] About half of the respondents were excluded from the present analysis because they did not provide information regarding participation in cultural activities and consumption.

[3] This idea of disaggregated occupational class seems to parallel what Yosano (1998) suggested.

The words in bold in parentheses are used to represent each occupational category, and abbreviation **l** and **s** stand for large and small, respectively.

To clarify the relationship between social stratification position, which is categorical data, and cultural consumption, multiple correspondence analysis (MCA) was applied. We used configuration space to map both the cultural consumption and occupational groups which allow us to identify cultural boundaries among social groups in terms of cultural consumption patterns (Greenacre 1984). We analyzed men and women separately because it is also of interest to look at gender differences in the structural relationship between social groups and lifestyles.

3 Results

Results for the analysis of male respondents are reported in Fig. 1 and for female respondents in Fig. 2.[4] The figure represents the positions of the social groups (with square markers) and the cultural consumption patterns (with circle markers) in two-dimensional space.

3.1 Males Cultural Consumption and Occupational Stratification

The points depicting the respondents' occupational status form a semi-circle, which is sometimes called a horseshoe curve, when viewed collectively. Therefore, this suggests that the first dimension (horizontal) represents the occupational hierarchical order in terms of cultural consumption patterns. Most prestigious groups that actively participate in the cultural activities highly evaluated are in the right half, whereas those with very limited participation in cultural activities are in the left half. And also, most of the active engagements are on the right-hand side, and many of the disengagements are on the left-hand side.

However, cultural consumption that is traditionally thought to be genteel or highbrow (e.g., attending classical music concerts, museums and art exhibitions) does not necessarily locate in the extreme right half. Instead, the activities based on information literacy (e.g., drawing upon Restaurant Guides and mail order catalog shopping) or new forms of consumption (e.g., credit card purchasing and online shopping and booking) locate at the extreme. Ganzeboom (1982) suggests that individual differences in information-processing capacity influence people's cultural consumption patterns. The finding here seems consistent with this hypothesis.

The status order of occupations along dimension 1 (horizontal axe) is also not systematically arrayed as might be assumed for men. Occupational group engaged in creating cultural products is associated with the most progressive cultural consumption and is found at the extreme prestigious end of the scale. At the same time,

[4] Not all leisure activity categories are shown in the figure. + or − represents frequency of engagement in various activities (high, low, never, etc.).

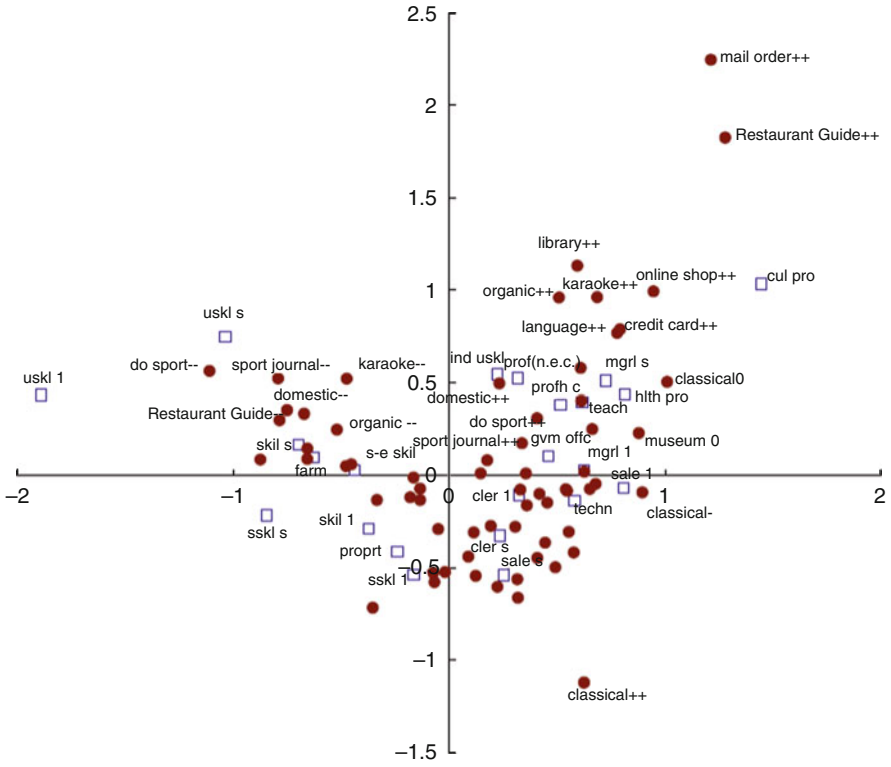


Fig. 1 Structure of male cultural consumption and social class: Dimension 1 × Dimension 2

however, they do not correspond closely to highbrow culture in the conventional sense (e.g., appreciation of classical music and visiting art exhibitions). They are, rather, characterized by rationality, efficiency, and a high degree of result calculability (e.g., mail order catalog shopping and dining out for pleasure drawing upon Restaurant Guides). The results here can be understood in the light of the argument by Erickson (1996), which is that the business elite are more familiar with consumption styles that appear to have more direct profitability for them and therefore they exclude high culture in favor of this lifestyle. Having said that, the finding that cultural and artistic producers are the most culturally active and possess cultural capital is concordant with the argument by Bourdieu.

The second highest group, somewhat below the highest level of occupational status, is the cluster of occupations that spend the most time talking to or interacting with others who are at various levels in the hierarchy, such as healthcare professionals, managers in a small firm, and sales in large firm. That is followed by the social group that is seen as privileged elite in our society (e.g., managers in large firms, professionals with higher educational credentials). This group tends to enjoy highbrow culture more than the so-called cultural elite (i.e., cultural producer), but at the same time they consume a sort of business culture, such as karaoke and language

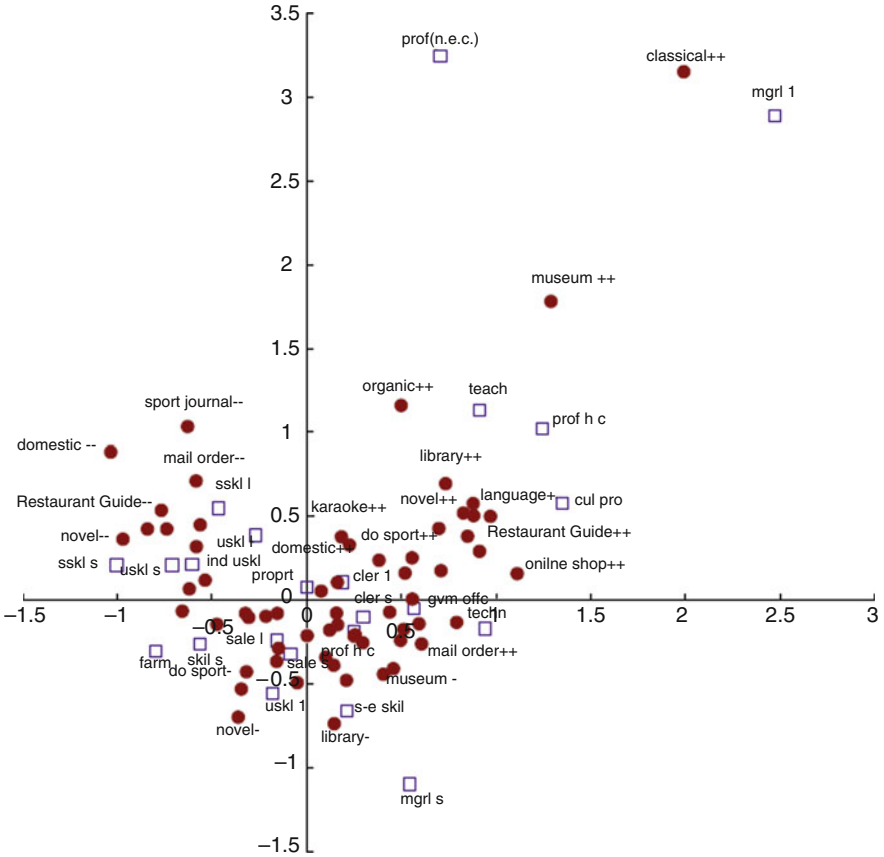


Fig. 2 Structure of female cultural consumption and social class: Dimension 1 × Dimension 2

study. The results also show that there is a gap between the official, or public, sector and the private sector amongst clerical workers in terms of cultural taste. Clerical workers in the public sector are clearly associated with middle-class cultural consumption bolstered by consciousness or literacy, but those in the private sector are only halfheartedly involved in those activities. Also, moderately high status people are characterized by active involvement in eco-conscious consumption, such as purchasing home-grown and organic food.

On the other side of the dimension, most blue-collar classes show up. Those who work in unskilled occupations in large organizations are the most culturally deprived or excluded, rather than those who work for small and medium-sized companies. They are excluded from most cultural practices and therefore have little cultural capital. This is probably due to the rise in non-regular unskilled male workers in large organizations over the past several years. Non-regular workers, especially those working on short-term contracts and as dispatched workers, receive little firm-based

training, low employment protection, and also low wages. This group is excluded from cultural life as a whole.

Dimension 2 (vertical) seems to be interpreted as the types of cultural activities. This roughly separates utilitarian or pragmatic in the upper half, and aesthetic in the lower half.

3.2 Female Cultural Consumption and Occupational Stratification

Figure 2 maps the co-ordinates for female respondents. The occupational categories are depicted as points which form a horseshoe curve, as were also the results for men. The managerial category is towards the top right, clerical and self-employed categories are in the lower central region to the origin, and the semi-skilled manual and unskilled work categories are towards left. Therefore the first dimension is an overall dimension which put the respondents and their cultural consumption activities in order from 'distinguished' on the right to 'uncultured and inactive' on the left, with groups on the right prestigious status and the lower ranking on the left.

As compared with the findings for men, though, the ordering found in the analysis of women are interpreted differently. This ordering is very much what one would expect and is not surprising, in contrast to the results for men. We can see that managerial occupations in large firms, which is a male-dominated highest status group, is at the extreme prestigious end of the scale, followed by professionals with higher educational credentials. Cultural and artistic producers, which is one of the prestigious occupational categories, is not seen to be as exclusive because they tend to have more access not only to highbrow culture but also to business culture.

The aesthetic disposition of the female managerial group seems to reflect a distinctive expression of a privileged position in social space. This suggests that highbrow culture is useful in female privileged groups for domination, but not for men. Highbrow culture might be seen as an irrelevant waste of time for the male business elite in Japan. Middle class women are also characterized by active involvement in eco-conscious consumption, such as purchasing domestic and organic food as are middle class men.

We find those who avoid most cultural activities to be located on left half. A little closer to the intersect we see routine non-manual employees and petty bourgeoisie; they have preferences for popular culture. Sales, proprietors and most blue-collar groups are neither cultured nor active.

4 Conclusion and Discussion

Based on our analysis which revised conventional class maps and included wider domains of consumption and lifestyle, we were able to confirm the inter-related nature of occupational class and lifestyle. In general, for both men and women, those

who are well-educated and those having higher occupational positions have more cultural capital. However, men and woman have different cultural specialties and patterns of correspondence with occupational class position. For women, highbrow culture is important for distinguishing themselves and maintaining their social position. For men of higher position, by contrast, highbrow culture is an irrelevant waste of time and instead business culture, which is characterized by a mixture of enterprise and rationality, prevails. Cultural capital in the contemporary world of business is based more on practical activities than on acquiring *established* accomplishments.

Further research must be undertaken in these areas that utilizes the data of a follow-up study. Due to the rapid global economic downturn, it is becoming ever more interesting to see how leisure practices are unequally allocated among people. Changes in economic and social structures have led to a continuous change in the relationship between people's social class and cultural consumption through the cumulative processes of social disadvantage.

Acknowledgements I thank the Social Stratification and Social Mobility (SSM) 2005 Committee for the use of data.

References

Bourdieu, P. (1979). *La Distinction: Critique Sociale du Jugement*. Paris: Minuit.

DiMaggio, P. (1987). Classification in art. *American Sociological Review*, *52*(4), 440–455.

Erickson, B. H. (1996). Culture, class, and connections. *American Journal of Sociology*, *102*(1), 217–251.

Ganzeboom, H. (1982). Explaining differential participation in high-cultural activities: A confrontation of information-processing and status seeking theories. In W. Raub (Ed.), *Theoretical models and empirical analyses* (pp. 186–205). Utrecht: E. S. -Publications.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Kataoka, E. (1992). Social and cultural reproduction process in Japan. *Sociological Theory and Methods*, *11*, 33–55.

Nakai, M. (2005). Cultural aspects and social stratification of women: Overlapping clustering and multidimensional scaling. *Proceedings of the 29th Annual Conference of the German Classification Society* (p. 246), German Classification Society (GfKl).

Nakai, M. (2008). Social stratification and gendered cultural consumption and lifestyles. In T. Sugano (Ed.), *Social class and disparities in quality of life* (pp. 1–28), SSM Research Series. Sendai, Japan: SSM2005 Committee.

Peterson, R. A., & Simkus, A. (1992). How musical tastes mark occupational status groups. In M. Lamont & M. Fournier (Eds.), *Cultivating differences. Symbolic boundaries and the making of inequality* (pp. 152–186). Chicago: The University of Chicago Press.

Veblen, T. B. (1899). *The theory of the leisure class: An economic study in the evolution of institutions*. Boston: Houghton Mifflin.

Weeden, K., & Grusky, D. B. (2005). The case for a new class map. *American Journal of Sociology*, *111*(1), 141–212.

Yosano, A. (1998). The changes of social class and the class openness. In A. Yosano (Ed.), *Social stratification and industrialization* (pp. 43–63), SSM Research Series. Sendai, Japan: SSM2005 Committee.

Centrality of Asymmetric Social Network: Singular Value Decomposition, Conjoint Measurement, and Asymmetric Multidimensional Scaling

Akinori Okada

Abstract The centrality of an asymmetric social network, where relationships among actors are asymmetric, is investigated by singular value decomposition, conjoint measurement, and asymmetric multidimensional scaling. They were applied to asymmetric relationships among managers of a small firm. Two sets of outward and inward centralities were derived by singular value decomposition. The first set is similar to the centralities obtained by conjoint measurement and by asymmetric multidimensional scaling. The second set represents the different aspect from the centralities of the first set as well as those derived by the conjoint measurement and the asymmetric multidimensional scaling.

1 Introduction

The centrality of a social network represents the relative importance, salience, power or attractiveness of an actor of a social network. The social network has been assumed to be symmetric, and characteristic values and vectors were used to derive the centrality (Bonacich 1972, Richards and Seary 2000). Bonacich (1972) used the characteristic vector corresponding to the largest characteristic value. Okada (2008) introduced a procedure using two (or more) characteristic vectors, which presents two (or more) centralities for each actor.

The relationships among actors can be asymmetric; relationship from actors j to k is not necessarily equal to the one from actors k to j (Wasserman and Faust 1994, pp. 510–511). Then we have an asymmetric social network (Barnett and Rice 1985, Bonacich and Lloyd 2001, Tyler et al. 2003). The asymmetric social network can be studied by singular value decomposition (Okada 2010; cf. Write and Evitts 1961), conjoint measurement, or asymmetric multidimensional scaling. Okada (2010) extended the earlier study (Okada 2008), to derive two sets of outward and inward centralities of the asymmetric social network by singular value decomposition. The present study is the extension of earlier studies (Okada 2008, 2010) based on singular value decomposition to derive the centrality of the asymmetric social network. The purpose of the present study is to compare the procedure based

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_25,

on singular value decomposition with additive conjoint measurement and asymmetric multidimensional scaling in dealing with an asymmetric social network, and to clarify the characteristics of the procedure based on singular value decomposition.

2 The Procedures

Three procedures; singular value decomposition, conjoint measurement, and asymmetric multidimensional scaling, were applied to the matrix of relationships among actors called $\mathbf{A}$, where a_{jk}, the (j,k) element of $\mathbf{A}$, represents the closeness of relationship from actors j to k. The singular vector of $\mathbf{A}$ gives the centrality. The singular value decomposition in the present study is characterized by deriving two sets of centralities (Okada 2008, 2010). The first set consists of the first left and right singular vectors corresponding to the largest singular value. The left singular vector represents the outward centrality which expresses the strength of the tendency of extending relationships from the corresponding actor to the other actors. The right singular vector represents the inward centrality which expresses the strength of the tendency of accepting relationships from the other actors to the corresponding actor along dimension 1. The second set consists of the second left and right singular vectors corresponding to the second largest singular value, and represents the outward and the inward centrality of an actor respectively along dimension 2.

Let $\mathbf{x}_i$ be the i-th left singular vector, and $\mathbf{y}_i$ be the i-th right singular vector, both corresponding to the i-th largest singular value ($i = 1$ or 2). The j-th element of $\mathbf{x}_i$, x_{ji}, represents the outward centrality of actor j, and the k-th element of $\mathbf{y}_i$, y_{ki}, represents the inward centrality of actor k along dimension i. The closeness of the relationship from actors j to k is represented by using the first and second left and right singular vectors and values. $\mathbf{A}$ is represented by

$$\mathbf{A} \cong \mathbf{X}_{12}\mathbf{D}\mathbf{Y}'_{12}, \tag{1}$$

where $\mathbf{X}_{12}$ is the matrix whose first and second columns are $\mathbf{x}_1$ and $\mathbf{x}_2$, $\mathbf{Y}_{12}$ is the matrix whose first and second columns are $\mathbf{y}_1$ and $\mathbf{y}_2$, and $\mathbf{D}$ is the diagonal matrix having the first and second singular values d_1 and d_2 as its diagonal elements. The closeness of the relationships from actors j to k is represented by

$$a_{jk} \cong d_1 x_{j1} y_{k1} + d_2 x_{j2} y_{k2}. \tag{2}$$

In dealing with $\mathbf{A}$ by additive conjoint measurement, we think $\mathbf{A}$ consists of two attributes. Rows of $\mathbf{A}$ are regarded as an attribute, and each row is regarded as a level of the attribute. Columns of $\mathbf{A}$ are regarded as another attribute, and each column is regarded as a level of the attribute. The additive conjoint measurement of $\mathbf{A}$ gives a partial utility to each row, and gives another partial utility to each column. The partial utility given to the j-th row, ur_j, represents the outward centrality of actor j,

and the partial utility given to the k-th column, uc_k, represents the inward centrality of actor k. The closeness of the relationship from actors j to k is represented by

$$a_{jk} \cong ur_j + uc_k. \tag{3}$$

The asymmetric multidimensional scaling (Okada and Imaizumi 1987) of **A** gives a configuration of actors. Each actor is represented as a point and a circle in a two-dimensional configuration (sphere or hypersphere in a three- or more than three-dimensional configurations) centered at that point in a multidimensional Euclidean space. Interpoint distances in the configuration represent symmetric relationships among actors, and the radius of the circle of an actor represents the relative asymmetry of the actor. The larger the radius of an actor is, the larger the outward centrality of that actor is and the smaller the inward centrality of that actor is. The radius represents the strength of the outward centrality and the weakness of the inward centrality of the corresponding actor as well.

3 The Data

In the present study relationships among 21 managers at a small firm, which were dealt with in earlier studies (Krackhardt 1987, Okada 2010), were analyzed. They consist of one president (#7), four vice presidents (#2, #14, #18, and #21), and 16 supervisors. Each vice president heads up a department, and each of the 16 supervisors belongs to one of the four departments (Table 1). The data consist of 21 matrices where each comes from each manager. The (j, k) element of the i-th matrix represents whether manager i perceived that managers j goes to k for advice at work; the (j, k) element $= 1$ when manager i perceived managers j goes to k for advice at work, and the (j, k) element $= 0$ otherwise.

The sum of 21 matrices, called matrix **A**, from 21 managers was analyzed. The (j, k) element of **A**, a_{jk}, represents the number of managers who perceived that managers j goes to k for advice at work, and shows the closeness perceived by 21 managers including manager j oneself. The maximum possible value, 21, is embedded in the diagonal of the matrix. The resulting matrix **A** shown in Table 2 is asymmetric, because a_{jk} is not necessarily equal to a_{kj}.

Table 1 Four departments and vice presidents

Department	Vice president	Supervisors
1	21	6, 8, 12, 17
2	14	3, 5, 9, 13,15 19, 20
3	18	10, 11
4	2	1, 4, 16

Table 2 The sum of 21 matrices

	To manager																				
From	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Manager 1	21	21	7	11	5	1	1	7	2	6	13	0	5	2	2	11	2	19	5	0	2
Manager 2	10	21	7	11	5	8	21	9	2	4	6	7	4	14	2	7	6	15	5	0	15
Manager 3	8	16	21	2	2	3	12	2	3	8	11	2	0	20	1	2	3	14	2	2	6
Manager 4	9	18	0	21	0	7	2	15	0	6	5	6	0	0	0	5	2	11	0	1	11
Manager 5	7	14	4	1	21	3	12	2	2	4	12	2	4	20	2	4	6	16	10	10	8
Manager 6	2	11	1	7	2	21	15	4	2	1	2	11	1	7	2	2	9	1	2	2	21
Manager 7	2	17	6	2	3	8	21	3	3	2	11	3	2	19	3	2	6	13	3	6	20
Manager 8	7	16	4	11	0	6	10	21	1	7	10	5	0	2	0	4	5	14	0	2	18
Manager 9	5	11	7	0	11	5	8	2	21	3	5	3	6	21	6	3	5	12	11	6	6
Manager 10	3	8	7	3	4	1	8	8	1	21	8	0	4	3	2	9	2	20	4	6	0
Manager 11	9	15	6	3	6	1	20	7	1	3	21	0	5	11	4	1	1	19	5	2	1
Manager 12	1	8	0	10	0	15	8	7	0	0	0	21	0	4	1	0	7	3	0	1	20
Manager 13	9	11	4	0	17	2	7	1	11	4	11	0	21	21	5	4	2	16	15	8	1
Manager 14	2	16	9	1	10	5	20	4	11	4	9	4	7	21	9	1	4	17	10	10	17
Manager 15	4	14	2	1	8	4	10	2	6	2	7	1	3	21	21	4	3	17	11	12	5
Manager 16	13	19	2	11	0	3	0	8	0	11	3	0	0	3	0	21	2	15	0	3	1
Manager 17	6	15	5	4	5	15	17	6	2	1	3	12	5	9	3	2	21	6	5	3	21
Manager 18	9	17	9	5	5	5	19	8	3	11	15	3	5	16	5	8	5	21	5	8	10
Manager 19	8	14	9	0	13	2	9	2	4	7	12	0	4	21	7	3	5	16	21	11	2
Manager 20	6	15	6	5	9	8	11	5	5	8	10	5	6	21	9	8	4	16	8	21	6
Manager 21	0	16	4	4	0	11	21	6	0	0	0	10	0	15	0	0	10	6	0	3	21

4 The Analysis and the Results

Five largest singular vales, derived by the singular value decomposition of **A** were 169.4, 66.5, 55.2, 30.7, and 22.9. In the present study, two sets of outward and inward centralities of 21 managers corresponding to the two largest singular values were derived. The first set based on the largest singular value is shown in Fig. 1; the outward centrality along dimension 1 (abscissa) which is the first left singular vector, and the inward centrality along dimension 1 (ordinate) which is the first right singular vector. The second set based on the second largest singular value is shown in Fig. 2; the outward centrality along dimension 2 (abscissa) which is the second left singular vector and the inward centrality along dimension 2 (ordinate) which is the second right singular vector.

The conjoint measurement of **A** resulted in the minimized stress of 0.632. Diagonal elements of **A** are regarded as missing values in the analysis. The partial utility given to the row represents the outward centrality of an actor, and the partial utility given to the column represents the inward centrality of an actor. They are represented in Fig. 3.

The asymmetric multidimensional scaling of **A** represents the minimized stress from five- through unidimensional spaces; 0.426, 0.426, 0.478, 0.538, and 0.792. Diagonal elements of **A** are not dealt with in the analysis. The two-dimensional

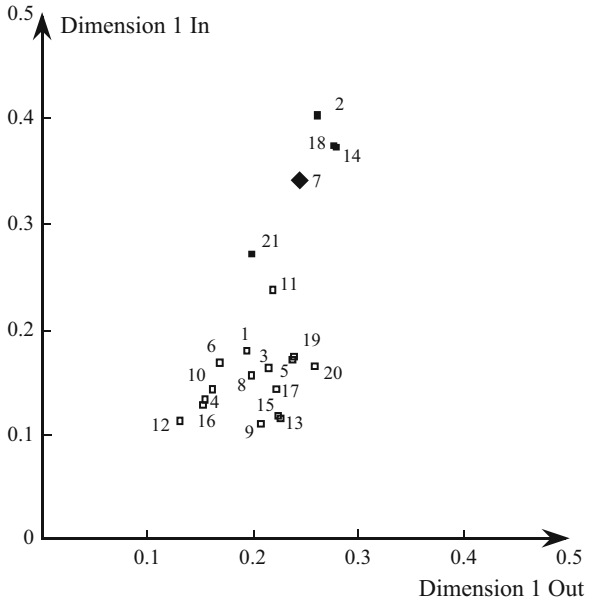


Fig. 1 The outward centrality based on the first left singular vector, and the inward centrality based on the first right singular vector. In Figs. 1–4, the president (#7) is represented by a *larger solid rhombus*, the vice president (#2, #14, #18, and #21) is represented by a *smaller solid square*, and each of the other supervisors is represented by an *open square*

result was chosen as the solution which is shown in Fig. 4. In the obtained two-dimensional configuration, manager j is represented as a point and a circle of radius r_j centered at the point corresponding to manager j.

5 Discussion

In Fig. 1, five upper ranked managers; the president (#7) and four vice presidents (#2, #14, #18, and #21), grouped together in the upper area of the configuration, and 16 supervisors or lower ranked managers grouped together in the lower area. While two groups have similar magnitude of the outward centrality, the upper ranked managers have the larger inward centrality than that the lower ranked managers have. This suggests that the upper ranked managers tend to be talked to by the other managers more frequently than the lower ranked managers, while all managers tend to talk to the other managers similarly.

While in Fig. 1 all managers are in the first quadrant, in Fig. 2 they are mainly in the first and the third quadrants. In Fig. 1 the product of the outward and inward centralities along dimension 1 is positive, or the relationship is positive among all managers. In Fig. 2 not all products are positive. When manager j's outward

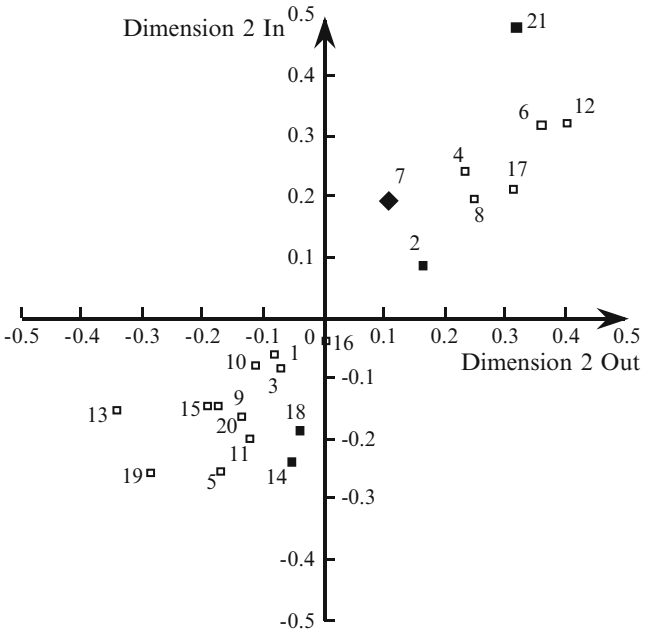


Fig. 2 The outward centrality based on the second left singular vector, and the inward centrality based on the second right singular vector

centrality is positive and managers k's inward centrality is negative, the relationship (the product of the outward and inward centralities) from managers j to k is negative along dimension 2, and the positive relationship (product) along dimension 1 is reduced as shown in (2). The president (#7) and two vice presidents (#2 and #21) have positive outward and inward centralities, but the other two vice presidents (#14 and #18) have negative outward and inward centralities. This means that the president (#7) and the two vice presidents (#2 and #21) have positive relationships with each other, and that the two vice presidents (#14 and #18) have positive relationships with each other as well. It is also suggested that the former (#7, #2, and #21) and the latter (#14 and #18) have negative relationships with each other. Manager #21 and all four managers belonging to Department 1 headed by Manager #21 are in the first quadrant, suggesting that all managers in Department 1 have positive relationships each other. This means that the department is cohesive, which is supported by that they are closely located at the right-hand area in Fig. 4.

The second set classifies managers (excluding manager #16) into two groups; one consists of managers in the first quadrant and the other consists of managers in the third quadrant. Two groups have positive relationships within each group which increase the positive relationships along dimension 1, and have negative relationships between two groups which decrease the positive relationships along dimension 1. It seems desirable to point out that two sets of outward and inward centralities given by the singular value decomposition represent different aspects of

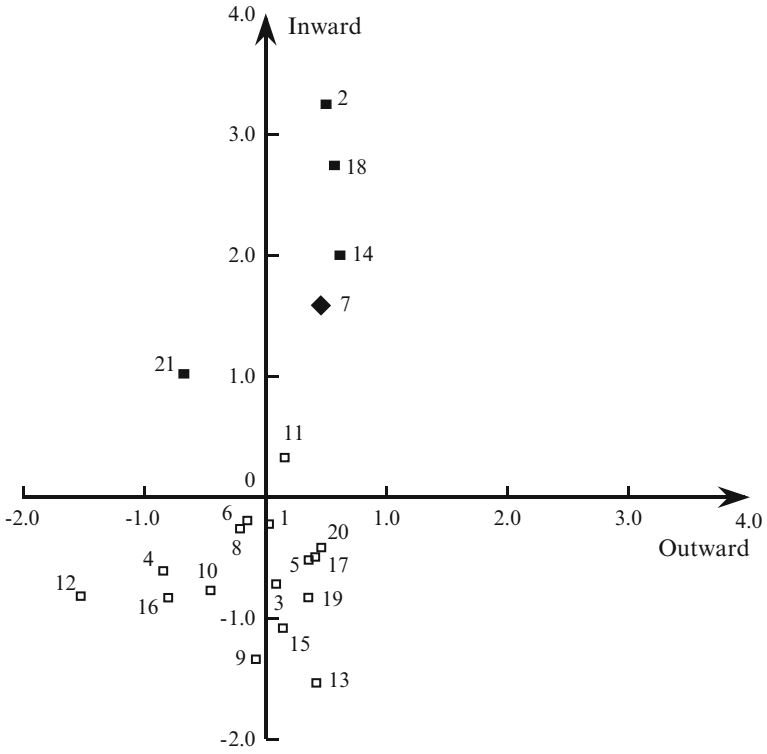


Fig. 3 The outward centrality and the inward centrality derived by conjoint measurement

centralities. The correlation coefficient between two sets of outward centralities is −0.44, and that of inward centralities is −0.02.

In Fig. 3, five higher ranked managers have higher inward centrality, while all managers have similar outward centrality. This coincides with the outward and inward centralities of the first set by the singular value decomposition (Fig. 1). The correlation coefficient between the outward centrality of the first set and the outward centrality of the conjoint measurement is 0.91, and that between the inward centralities is 0.98. While the manner of representing the relationships among managers is different (Eqs. 2 and 3), the two procedures derived similar results.

In Fig. 4 five upper ranked managers are located in the central part of the configuration. They have smaller radii than the other managers have. Manager #2, one of the vice presidents, has the smallest radius of zero by the definition (Okada and Imaizumi 1987, Eq. 2–4). This means that the upper ranked managers tend to be talked to more by the other managers than they tend to talk to the other managers. This is compatible with the outward and inward centralities of the first set by the singular value decomposition shown in Fig. 1. The correlation coefficient between the radius and the difference of outward and inward centralities of the first set is 0.92. The correlation coefficient between the radius and the difference of outward

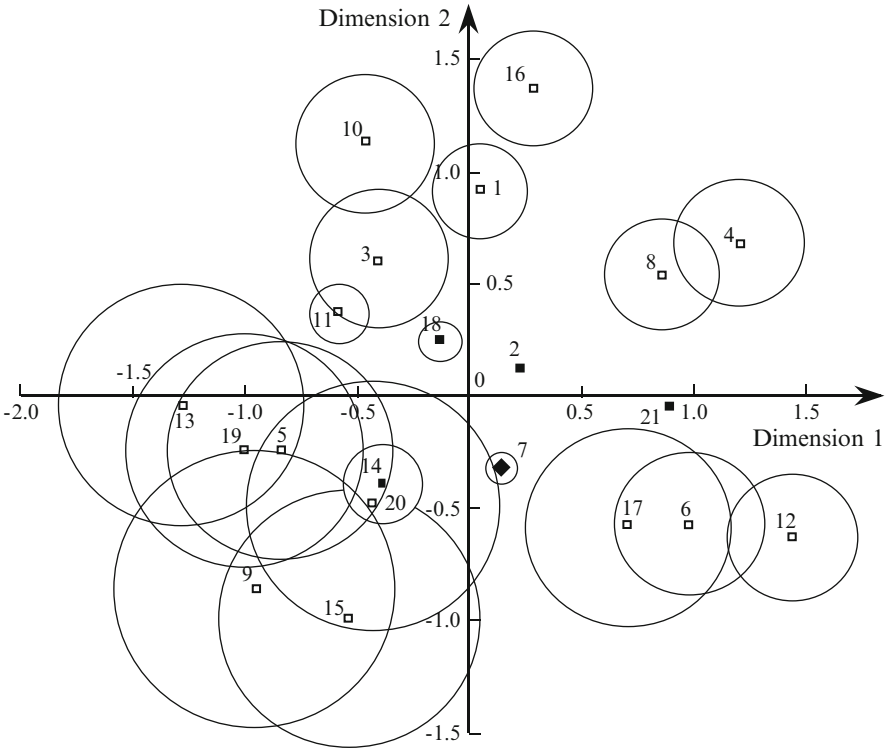


Fig. 4 The two-dimensional configuration derived by asymmetric multidimensional scaling

and inward centralities derived by the conjoint measurement is 0.90. These correlation coefficients suggest that the outward and inward centralities of the first set by the singular value decomposition, those derived by the conjoint measurement, and the radius derived by the asymmetric multidimensional scaling represent essentially the same aspect of the centrality of the relationships among managers.

The correlation coefficient between the outward centralities of the second set by the singular value decomposition and those derived by the conjoint measurement is -0.51, and that of the inward centralities is 0.09. The correlation coefficient between the radius derived by the asymmetric multidimensional scaling and the difference of outward and inward centralities of the second set is -0.22. The second set of inward and outward centralities represent different aspects from those of the first set as well as from those of the conjoint measurement and the asymmetric multidimensional scaling.

Acknowledgements The author wishes to express his appreciation for the benefit he received from discussions with Hiroshi Inoue and with Satoru Yokoyama. He also would like to express his gratitude to Antonella Plaia for her discussion and suggestions as a discussant given at the session of the 7th cladag meeting. He indebted to an anonymous referee for her/his valuable review on the earlier version of the paper, and to Reginald Williams for his help concerning English.

References

Barnett, G. A., & Rice, R. E. (1985). Longitudinal non-Euclidean networks: Applying GALILEO. *Social Networks*, *7*, 287–322.

Bonacich, P. (1972). Factoring and weighting approaches to clique identification. *Journal of Mathematical Sociology*, *2*, 113–120.

Bonacich, P., & Lloyd, P. (2001). Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, *23*, 191–201.

Krackhardt, D. (1987). Cognitive social structures. *Social Networks*, *9*, 109–134.

Okada, A. (2008). Two-dimensional centrality of a social network. In C. Preisach, L. Burkhardt, & L. Schmidt-Thieme (Eds.), *Data analysis, machiine learning and applications* (pp. 381–388). Heidelberg, Germany: Springer.

Okada, A. (2010). Two-dimensional centrality of asymmetric social network. In N. L. Lauro, et al. (Eds.), *Data analysis and classification* (pp. 93–100). Heidelberg, Germany: Springer.

Okada, A., & Imaizumi, T. (1987) Nonmetric multidimensional scaling of asymmetric similarities. *Behaviormetrika*, *21*, 81–96.

Richards, W., & Seary, A. (2000). Eigen analysis of networks. *Journal of Social Structure*, *1*(2), 1–17.

Tyler, J. R., Wilkinson, D. M., & Huberman, B. A. (2003). *E-mail as spectroscopy: Automated discovery of community structure within organization*. e-print condmat/0303264.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, UK: Cambridge University Press.

Write, B., & Evitts, M. S. (1961). Direct factor analysis in sociometry. *sociometry*, *24*, 82–98.

Part VI
Classification

Some Perspectives on Multivariate Outlier Detection

Andrea Cerioli, Anthony C. Atkinson, and Marco Riani

Abstract We provide a selective view of some key statistical concepts that underlie the different approaches to multivariate outlier detection. Our hope is that appreciation of these concepts will help to establish a unified and widely accepted framework for outlier detection.

1 Introduction

The identification of outliers is an important step of any analysis of multivariate data. In a multivariate setting, this task poses more challenging problems than in the simpler case of a single variable for at least three basic reasons:

- outlyingness should be judged with respect to several (possibly many) dimensions simultaneously;
- there is no natural ordering of multivariate data on which 'extremeness' of an observation can be ascertained;
- simple graphical diagnostic tools like the boxplot are difficult to construct in more than one or two dimensions (Zani et al. 1998).

It is thus not surprising that the systematic study of multivariate outliers has a long history in the statistical literature and has led to remarkably different points of view. See, e.g., Hadi et al. (2009) and Morgenthaler (2006) for recent reviews on robust methods and outlier detection. In these and other reviews, it is acknowledged that the concern for outliers or grossly wrong measurements is probably as old as the experimental approach to science. The earliest reported historical references usually date back to the seventeenth century, with the first precise specifications subsequently given by Gauss and Legendre. Perhaps less known is the fact that the same concern was also present in Ancient Greece more than 2,000 years ago, as reported by Thucydides in his *History of The Peloponnesian War* (III 20, 3–4).[1]

[1] According to Thucydides, in 428 B.C. the Plataeans, besieged by the Spartans, excluded extreme measurements when estimating the height of the walls that their enemies had built around the city. In this way, they managed to break the siege.

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_26,

In the modern statistical era, until the early 1990s, alternative methods were developed following three essentially distinct streams of research, well documented in the classical book by Barnett and Lewis (1994, Chap. 7): robust techniques for multivariate outlier accommodation; formal tests of hypotheses for precise outlier identification and, thirdly, less formal diagnostic tools for exploratory analysis including intuitive inspection of the data. With a bit of humour, the supporters of these alternative schools of outlier methodology were sometimes called the *robustniks*, the *testniks* and the *diagnostniks*, respectively. The reconciliation of different 'outlier philosophies' was already seen as an ideal by Rousseeuw and Zomeren (1990) in their rejoinder twenty years ago, but it has still to be reached. Sect. 4 provides a suggestion in that direction.

It is not the goal of this paper to provide a comprehensive overview of the wealth of methods developed for the purpose of multivariate outlier identification. Rather, the idea is to guide the reader through a few key statistical concepts that underlie the different approaches and to see how they evolved over the years. Our hope is that appreciation of these concepts will help us in establishing a unified and widely accepted framework for outlier detection.

2 Outlier Detection and Testing

In a seminal paper Wilks (1963) laid down the statistical foundations of multivariate outlier detection. Let $y = (y_1, \ldots, y_n)'$ be a sample of v-dimensional observations from $N(\mu, \Sigma)$. The sample mean is $\hat{\mu}$ and the unbiased sample estimate of Σ is $\hat{\Sigma}$. Wilks derived the exact distribution of the n scatter ratios,

$$R_i = |(n-2)\hat{\Sigma}_{\{i\}}||(n-1)\hat{\Sigma}|^{-1} \qquad i = 1, \ldots, n,$$

where $\hat{\Sigma}_{\{i\}}$ is the unbiased estimate of Σ computed after deleting y_i. It is easily seen (Atkinson et al. 2004, pp. 44–46) that R_i is inversely related to the squared Mahalanobis distance of observation y_i

$$d_i^2 = (y_i - \hat{\mu})' \hat{\Sigma}^{-1} (y_i - \hat{\mu}), \tag{1}$$

so that the distributional results for R_i hold for d_i^2 as well. In particular,

$$d_i^2 = \frac{(n-1)^2}{n}(1 - R_i) \sim \frac{(n-1)^2}{n} \text{Beta}\left(\frac{v}{2}, \frac{n-v-1}{2}\right) \qquad i = 1, \ldots, n. \tag{2}$$

Wilks also showed how the smallest ratio $R_{(1)}$, or equivalently the largest squared distance $d_{(n)}^2$, can be used to test the outlyingness of the corresponding observation. This multivariate outlier detection rule focuses on the *intersection hypothesis* that no outlier is present in the data

$$H_{0s} : \{y_1 \sim N(\mu, \Sigma)\} \cap \{y_2 \sim N(\mu, \Sigma)\} \cap \ldots \cap \{y_n \sim N(\mu, \Sigma)\}, \tag{3}$$

against the alternative that one is present. The candidate outlier is the most remote observation, i.e., the observation with the largest squared Mahalanobis distance (1). The size of a test of H_{0s}, say γ, represents the proportion of good *data sets* that are wrongly declared to contain outliers. Simultaneity is dealt with in Wilks (1963) by introducing a Bonferroni bound on the probability that the test statistic exceeds a given threshold.

Simulations show that Wilks' outlier detection method, combining the scaled Beta distribution (2) with a Bonferroni bound, has very good control of the size of the test of H_{0s}. It can thus be taken as a benchmark for comparison with alternative procedures under the null hypothesis of no outliers.

However, the squared Mahalanobis distances (1) suffer from *masking*. If a few outliers contaminate the data, it is unlikely that the largest distances $d^2_{(n)}, d^2_{(n-1)}, \ldots$ will be associated with the atypical observations because $\hat{\mu}$ and $\hat{\Sigma}$ will be grossly distorted by these outliers.

Wilks (1963) extended his deletion method to the case of two observations, but dealing with an unknown and possibly large number of outliers rapidly becomes infeasible. The same problem affects any other backward procedure, such as the sequential application of Wilks' test suggested by Caroni and Prescott (1992). Moving backwards, all the outliers will be missed if $d^2_{(n)}$ is masked.

3 Robust Distances from High-Breakdown Estimators

The use of high-breakdown estimators of μ and Σ in the place of the classical ones has proved to be a practical solution to the problem of masking. Popular choices for such estimators include the Minimum Covariance Determinant estimator, S-estimators and projection-based techniques. See Hubert et al. (2008) or Maronna et al. (2006, Chap. 6) for recent reviews.

Let $\tilde{\mu}$ and $\tilde{\Sigma}$ be the chosen high-breakdown estimators of μ and Σ. The corresponding squared robust Mahalanobis distances are

$$\tilde{d}_i^2 = (y_i - \tilde{\mu})' \tilde{\Sigma}^{-1} (y_i - \tilde{\mu}) \qquad i = 1, \ldots, n. \tag{4}$$

The outliers in y are revealed by their large distances from the robust fit provided by $\tilde{\mu}$ and $\tilde{\Sigma}$, without suffering from masking. The key issue is that outlying observations have null or negligible weight in the computation of $\tilde{\mu}$ and $\tilde{\Sigma}$. Therefore, they cannot 'attract' these estimates and maintain a large value of $\tilde{d}_i^2$.

The need to avoid masking directs the outlier detection problem to the choice of suitable cut-offs for the robust squared distances (4), instead of the classical critical values computed from (2). However simple this step may seem, it has produced some surprising consequences.

First, the focus has shifted from the intersection hypothesis (3) to the problem of testing the n null hypotheses

$$H_{0i} : y_i \sim N(\mu, \Sigma), \qquad i = 1, \ldots, n. \tag{5}$$

The most common approach has been to test all these hypotheses individually at a specified size $0.01 \leq \alpha \leq 0.05$, with $\alpha = 0.025$ being perhaps the most popular choice. This approach, which does not take multiplicity of tests into account, increases the probability of detecting truly contaminated observations, but the user must be prepared to declare at least one outlier (and often many more) in most data sets of realistic size. In other words, the user must be prepared to invest a large sum of money (if all the suspected outliers are discarded) or a large amount of time (if the suspected outliers are checked one by one) to accomplish the process of multivariate outlier detection, even when the expected number of contaminated observations is small.

The tendency of an outlier detection method to label good observations as outliers is called *swamping*. We believe that the ability to control the degree of swamping is an important property for the practical usefulness of an outlier detection method. Major application areas where even a moderate number of false outliers may have disastrous consequences include anti-fraud analysis and statistical quality control. For instance, outliers are of great interest in the analysis of trade data arising in the European Union market (Riani et al. 2008), because some of them may correspond to fraudulent transactions. Since there are hundreds of transactions to be inspected over thousands of markets, ignoring the multiplicity of tests would lead to a plethora of false signals for anti-fraud services, thus making substantial investigation of possible frauds impractical.

Another major shortcoming of the use of the squared robust distances $\tilde{d}_i^2$ is that their exact distribution is unknown. The required cut-offs are then usually computed from their asymptotic χ_v^2 distribution, although the adequacy of this approximation can be very poor even in moderately large samples, especially when the number of dimensions increases. This behaviour has been shown in many simulation studies: see, e.g., Becker and Gather (2001), Cerioli et al. (2009), Hardin and Rocke (2005) and Riani et al. (2009).

The liberality of the χ_v^2 distribution for the purpose of approximating the squared robust distances $\tilde{d}_i^2$ adds further swamping to the individual testing framework of the n hypotheses H_{0i}. It also makes the simultaneous testing of these hypotheses in (3) even more problematic, because the corresponding cut-offs lie in the extreme tail of the true but unknown distribution of the robust distances. This behaviour is in sharp contrast with the excellent null performance of the classical Mahalanobis distances (1). It also obviously calls for better approximations to the finite sample distribution of the squared robust distances $\tilde{d}_i^2$ when no outlier is present in the data.

Hardin and Rocke (2005) suggest a way to approximate the first two moments of the distribution of the squared robust distances $\tilde{d}_i^2$. However, simultaneous testing of the n hypothesis (5) requires cut-off values which are in the extreme tail of the distribution. In that case information on $E(\tilde{d}_i^2)$ and $var(\tilde{d}_i^2)$ is often not enough to obtain reliable rejection regions and it is preferable to estimate the cut-offs directly. A good finite sample approximation to the required thresholds under the intersection hypothesis (3) is proposed by Cerioli et al. (2009). Their idea is to calibrate the asymptotic cut-off values by Monte Carlo simulation. Calibration is first performed for some representative values of n and v and then extended to any n and v by

parametric non-linear interpolation. The resulting outlier detection rule has very good control of the size of the test of no outliers even in situations where space is very sparsely filled (e.g., $n = 50, v = 10$). The method is very general, as in principle it can be applied to any choice of $\tilde{\mu}$ and $\tilde{\Sigma}$, and also easy to implement, once the parameters of the interpolation function are made available. However, the power may be rather low, since the technique does not allow for the variability of distances in the tail of the distribution.

Cerioli (2010) provides a power improvement by introducing an accurate approximation to the distribution of one-step reweighted robust distances. This approximation is based on a scaled Beta distribution mimicking (2) for the units not suspected of being outliers, and on a scaled F distribution for the units which are trimmed in the reweighting step. Also this method provides good control of the simultaneous size of the n outlier tests (5). Therefore, it can be useful in all the application fields where allowing for the multiplicity of tests is an important issue.

Attaining the right size through distributional results yields more powerful rules than through calibration of cut-off values. Furthermore, a substantial increase in power can be obtained by controlling the number of false discoveries only when all the data come from the prescribed null distribution. Cerioli (2010) suggests an outlier identification rule that tolerates some degree of swamping, but only when there is strong evidence that some contamination is present in the data. This follows the idea that the level of swamping provided by repeated testing of (5), although deleterious in 'good' data sets, may still be acceptable in a contaminated framework. Such a view is often sensible when the probability of observing a contaminated sample is small. On the contrary, if the sample is predicted to have some contamination with high probability, but the expected number of contaminants is small, other approaches could be followed. As shown in Cerioli and Farcomeni (2011), by controlling the False Discovery Rate it is possible to develop outlier identification rules for which the acceptable number of false discoveries depends explicitly on the number of outliers found.

4 Is a Reconciliation Possible?

Wilks' outlier test and the high-breakdown identification rules described by Hubert et al. (2008) have opposite attitudes towards two basic issues of multivariate outlier detection: the null hypothesis to be tested in order to label an observation as an outlier and the approach towards the control of the number of false discoveries. A reconciliation of these alternative philosophies can be found in the Forward Search method of Atkinson et al. (2004, 2010).

The basic idea of the Forward Search (FS) is to start from a small, robustly chosen, subset of the data and to fit subsets of increasing size, in such a way that outliers and other observations not following the general structure are clearly revealed by diagnostic monitoring. Let m_0 be the size of the starting subset. Usually $m_0 = v + 1$ or slightly larger. Let $S_*^{(m)}$ denote the subset of data fitted by the FS at

step m ($m = m_0, \ldots, n$). At that step, outlyingness of each observation y_i can be evaluated through the squared distance

$$d^2_{i*}(m) = \{y_i - \hat{\mu}_*(m)\}' \hat{\Sigma}_*(m)^{-1} \{y_i - \hat{\mu}_*(m)\}, \tag{6}$$

where $\hat{\mu}_*(m)$ and $\hat{\Sigma}_*(m)$ are the estimates of μ and Σ computed from $S^{(m)}_*$. The squared distances $d^2_{1*}(m), \ldots, d^2_{n*}(m)$ are then ordered to obtain the fitting subset at step $m+1$. Usually one observation enters the subset at each step, but sometimes two or more, when one or more then leave. Such occurrences are indicative of changes in structure or of clusters of outliers entering the subset.

Whilst $S^{(m)}_*$ remains outlier free, the squared distances $d^2_{i*}(m)$ will not suffer from masking and swamping. Therefore, they are a robust version of the classical Mahalanobis distance d^2_i. The main diagnostic quantity computed from these robust distances is $d^2_{i_{\min}*}(m)$, where

$$i_{\min} = \arg\ \min d^2_{i*}(m) \qquad i \notin S^{(m)}_*$$

is the observation with the minimum squared Mahalanobis distance among those not in $S^{(m)}_*$. The main idea is that the distance of the closest observation entering the subset at step $m+1$ will be large if this observation is an outlier. Its peculiarity will be clearly revealed by a peak in the forward plot of $d^2_{i*}(m)$.

The early developments of the FS aimed essentially to provide powerful plots for investigating the structure of regression and multivariate data, using quantities such as $d^2_{i_{\min}*}(m)$. Therefore, the FS might be seen a contribution of the 'diagnostic' school of outlier detection. However, it is paramount that any diagnostic quantity can result in a formal test if its null distribution is known and appropriate thresholds can be defined. The statistic $d^2_{i_{\min}*}(m)$ can be treated as a squared deletion distance on $m-1$ degrees of freedom, whose distribution is (Atkinson et al. 2004, pp. 43–44)

$$\frac{(m^2-1)v}{m(m-v)} F_{v,m-v}, \tag{7}$$

while $S^{(m)}_*$ remains outlier free. This statistic is based on $\hat{\Sigma}_*(m)$, which is a biased estimate of Σ, being calculated from the m observations in the subset that have been chosen as having the m smallest distances. As a result, Riani et al. (2009) propose a formal outlier test based on the FS by making use of the envelopes

$$V_{m,\alpha}/\sigma_T(m), \tag{8}$$

where $V_{m,\alpha}$ is the $100\alpha\%$ cut-off point of the $(m+1)$th order statistic from the scaled F distribution (7) and the factor

$$\sigma_T(m)^{-1} = \frac{m/n}{P(X^2_{v+2} < \chi^2_{v,m/n})} \tag{9}$$

allows for trimming of the $n-m$ largest distances. In (9) $\chi^2_{v,m/n}$ is the m/n quantile of χ^2_v and $X^2_{v+2} \sim \chi^2_{v+2}$.

The FS test for multivariate outlier identification based on thresholds derived from (8) does not require computation of the high-breakdown estimators $\tilde{\mu}$ and $\tilde{\Sigma}$. Furthermore, like the methods described in Sect. 2, this is a simultaneous test which has good control of the size of the test of no outliers (3). This property is made possible by the use of accurate finite sample distributional results for the squared Mahalanobis distances computed along the search. Nevertheless, the FS test does not suffer from masking, because it is the algorithm itself which is robust. Thus the FS test can cope with the same contamination rate as the high-breakdown methods sketched in Sect. 3.

We conclude that the Forward Search can provide a reconciliation of the three classical approaches to outlier detection introduced in Sect. 1. Being based on a flexible strategy in which the proportion of trimming is determined by the data, it enjoys high power. A further bonus of the Forward Search is its suitability for being easily adapted to cope with many different methodologies, including other multivariate techniques, linear and non-linear regression and correlated data modelling. For instance, by allowing a level of trimming smaller than 0.5, we believe that the Forward Search has the greatest potential among robust techniques to become a comprehensive approach through which cluster analysis and outlier detection could be performed under the same umbrella.

Acknowledgements The Authors are grateful to Spyros Arsenis and Domenico Perrotta for pointing out the reference to Thucydides.

References

Atkinson, A. C., Riani, M., & Cerioli, A. (2004). *Exploring multivariate data with the forward search*. New York: Springer–Verlag.

Atkinson, A. C., Riani, M., & Cerioli, A. (2010). The forward search: Theory and data analysis (with discussion). *Journal of the Korean Statistical Society*, *39*, 117–134.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). New York: Wiley.

Becker, C., & Gather, U. (2001). The largest nonidentifiable outlier: A comparison of multivariate simultaneous outlier identification rules. *Computational Statistics and Data Analysis*, *36*, 119–127.

Caroni, C., & Prescott, P. (1992). Sequential application of Wilks's multivariate outlier test. *Applied Statistics*, *41*, 355–364.

Cerioli, A. (2010). Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association*, *105*(489), 147–156.

Cerioli A., & Farcomeni, A. (2011). Error rates for multivariate outlier detection, Computational Statistics and Data Analysis, 55, 544–553.

Cerioli, A., Riani, M., & Atkinson, A. C. (2009). Controlling the size of multivariate outlier tests with the MCD estimator of scatter. *Statistics and Computing*, *19*, 341–353.

Hadi, A. S., Rahmatullah Imon, A. H. M., & Werner, M. (2009). Detection of outliers. *WIREs Computational Statistics*, *1*, 57–70.

Hardin, J., & Rocke, D. M. (2005). The distribution of robust distances. *Journal of Computational and Graphical Statistics*, *14*, 910–927.

Hubert, M., Rousseeuw, P. J., & Van Aelst, S. (2008). High-breakdown robust multivariate methods. *Statistical Science*, *23*, 92–119.

Maronna, R. A., Martin, D. R., & Yohai, V. J. (2006). *Robust statistics: Theory and methods*. New York: Wiley.

Morgenthaler, S. (2006). A survey of robust statistics. *Statistical Methods and Applications*, *15*, 271–293 (Erratum *16*, 171–172).

Riani, M., Cerioli, A., Atkinson, A., Perrotta, D., & Torti, F. (2008). Fitting mixtures of regression lines with the forward search. In F. Fogelman-Soulié, D. Perrotta, J. Piskorski, & R. Steinberger (Eds.), *Mining massive data sets for security* (pp. 271–286). Amsterdam: IOS Press.

Riani, M., Atkinson, A. C., & Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B*, *71*, 447–466.

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. With discussion. *Journal of the American Statistical Association*, *85*, 633–651.

Wilks, S. S. (1963). Multivariate statistical outliers. *Sankhya A*, *25*, 407–426.

Zani, S., Riani, M., & Corbellini, A. (1998). Robust bivariate boxplots and multiple outlier detection. *Computational Statistics and Data Analysis*, *28*, 257–270.

Spatial Clustering of Multivariate Data Using Weighted MAX-SAT

Silvia Liverani and Alessandra Petrucci

Abstract Incorporating geographical constraints is one of the main challenges of spatial clustering. In this paper we propose a new algorithm for clustering of spatial data using a conjugate Bayesian model and weighted MAX-SAT solvers. The fast and flexible Bayesian model is used to score promising partitions of the data. However, the partition space is huge and it cannot be fully searched, so here we propose an algorithm that naturally incorporates the geographical constraints to guide the search over the space of partitions. We illustrate our proposed method on a simulated dataset of social indexes.

1 Introduction

Clustering of spatial data requires an appropriate inclusion of the geographical distances in the clustering algorithm. One of the proposed ways to include the spatial element suggests to run clustering only between observations and clusters that share a boundary (Johnston 1976, Legendre 1987, Gordon 1996). Such an exploration of the space, though, does not automatically generate a partition of the data. Moreover, the huge number of partitions of even a moderately sized dataset requires an intelligent search algorithm.

In this paper we propose the use of a Bayesian score model, which allows the search over adjacent clusters, and a weighted MAX-SAT solver, that searches the space of high scoring clusters identifying the highest scoring partition and encodes the constraints necessary to obtain a partition of the data.

The Bayesian MAP model provides us with a score for each partition as in Heard et al. (2006). MAP selection, an alternative to model mixing (Fraley and Raftery 1998), is a method where the most a posteriori probable model is selected. As shown in the context of microarray experiments (Heard et al. 2006, Zhou et al. 2006, Liverani et al. 2009a), Bayesian algorithms are very versatile, for example to model time dependence and to enable incorporation of pertinent scientific information.

We propose the use of this Bayesian score in conjunction with a weighted MAX-SAT solver (Tompkins and Hoos 2005). Given a set of weighted clauses,

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_27,

the weighted maximum satisfiability problem (weighted MAX-SAT) asks for the maximum weight which can be satisfied by any assignment. Many exact solvers for MAX-SAT have been developed in recent years and it has been demonstrated that these solvers can be used in other fields too, when the problem can be encoded appropriately. For example, see Cussens (2008) for model search over Bayesian networks. In particular, Liverani et al. (2010) have successfully encoded the clustering problem to explore the partition space using weighted MAX-SAT solvers when a Crowley (1997) prior is employed.

The main contribution of this paper is a modification to the encoding by Liverani et al. (2010) to the context of spatial clustering. Some of the difficulties of the implementation of the above algorithm to a generic clustering problem are immediately simplified in this context by encoding geographical constraints when scoring the promising clusters searched by weighted MAX-SAT solvers. Moreover, in this context weighted MAX-SAT solvers become essential because the geographical constraints do not allow the use of alternative algorithms. For example, agglomerative hierarchical clustering (AHC) would obtain a partition whose data points might not share a boundary. However, even though weighted MAX-SAT solvers identify partitions given a set of potential clusters, they do require a careful computation of those potential clusters, including the geographical constraints. In this paper we also include an algorithm for selecting such clusters.

This paper is structured as follows. In Sect. 2 we introduce the Bayesian scoring function. In Sect. 3 we introduce weighted MAX-SAT solvers and the encoding of geographical constraints. In Sect. 4 we illustrate our method on a simulated dataset.

2 Bayesian Model and Score Function

We perform clustering with a conjugate Gaussian regression model developed by Heard et al. (2006).

Let $\boldsymbol{Y}_i \in \mathbf{R}^r$ for $i = 1, \ldots, n$ represent the r-dimensional units to cluster. Let $D = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ and $\boldsymbol{Y} = \text{vec}(D)$. The Normal Inverse-Gamma conjugate Bayesian linear regression model for each observation in a cluster c takes the form

$$Y^{(c)} = X^{(c)}\boldsymbol{\beta}^{(c)} + \varepsilon^{(c)},$$

where $\boldsymbol{\beta}^{(c)} = (\beta_1, \beta_2, \ldots, \beta_p)' \in \mathbf{R}^p$ and

$$\varepsilon^{(c)} \sim N(\mathbf{0}, \sigma^2_{(c)} I)$$

is a vector of independent error terms with $\sigma^2_{(c)} > 0$. The posterior Normal Inverse Gamma joint density of the parameters $(\boldsymbol{\beta}^{(c)}, \sigma^2_{(c)})$ is given in Heard et al. (2006). A partition $\mathscr{C}$ of the observations divides them into N clusters of sizes $\{n_1, \ldots, n_N\}$, with

$$n = \sum_{i=1}^{N} n_i .$$

Throughout this paper we assume that $X = \mathbf{1}_{n_c} \otimes B$, where B is a known matrix, and that $X'X = n_c B'B$ is full rank. The design or basis function matrix X encodes the type of basis used for the clustering: for example, linear splines (Heard et al. 2006), wavelets (Ray and Mallick 2006) and Fourier (Liverani et al. 2009a). For merely illustrative purposes, in Sect. 4 we use a design matrix which represents a linear relation between the coefficients $\boldsymbol{\beta}$ and the response variable $\boldsymbol{Y}$.

The Bayes factor associated with this model can be calculated from its marginal likelihood, $L(\boldsymbol{y})$ (Heard et al. 2006), that is,

$$BF = \left(\frac{1}{\pi}\right)^{nr/2} \frac{b^a}{(b^*)^{a^*}} \frac{|V^*|^{1/2}}{|V|^{1/2}} \frac{\Gamma(a^*)}{\Gamma(a)}. \tag{1}$$

Unlike for univariate data, there are a myriad of different shapes given by multivariate data and, consequently, Bayes factors associated with different observations are highly discriminative and informative.

Assuming the parameters of different clusters are independent then, because the likelihood separates, it is straightforward to check (Smith et al. 2008) that the log marginal likelihood score $\Sigma(\mathscr{C})$ for any partition $\mathscr{C}$ with clusters $c \in \mathscr{C}$ is given by

$$\Sigma(\mathscr{C}) = \log p(\mathscr{C}) + \sum_{c \in \mathscr{C}} \log p_c(\boldsymbol{y}).$$

This model has a wide applicability because it can be customized through the choice of a given design matrix X. Conjugacy ensures the fast computation of scores for a given partition because these can be written explicitly and in closed form as functions of the data and the chosen values of the hyperparameters of the prior. Applications range from one-dimensional data points to multidimensional datasets with time dependence among points or where the points are obtained by applying different treatments to the units. No geographical constraints are included in the model at this stage.

Although there are many possible choices for a prior over partitions, an appropriate choice in our context is the Crowley partition prior $p(\mathscr{C})$ for partition $\mathscr{C}$ (Liverani et al. 2010). This prior is given by

$$p(\mathscr{C}) = \frac{\Gamma(\lambda)\lambda^N}{\Gamma(n+\lambda)} \prod_{i=1}^{N} \Gamma(n_i)$$

where $\lambda > 0$ is the parameter of the partition prior, N is the number of clusters and n is the total number of observations, with n_i the number of observations in cluster c_i.

Assuming the parameters of different clusters are independent then, because the likelihood separates (Smith et al. 2008), if we use a prior from this family then we can compute the score $\Sigma(\mathscr{C})$, which decomposes into the sum of the scores S_i over individual clusters plus a constant term. Thus,

$$\Sigma(\mathscr{C}) = \log \Gamma(\lambda) - \log \Gamma(n + \lambda) + \sum_{i=1}^{N} S_i.$$

This is especially useful for weighted MAX-SAT which needs the score of an object to be expressible as a sum of component scores. It is this property that enables us to find straightforward encoding of the MAP search as a weighted MAX-SAT problem.

3 Searching the Partition Space

The Bayesian score introduced above is used in conjunction with a method for the search of the partition space because a full exploration of the partition space is not possible, as this space is huge. The number of partitions of a set of n elements grows quickly with n. For example, there are 5.1×10^{13} ways to partition 20 elements.

Liverani et al. (2010) showed that a decomposition of the marginal likelihood score allows weighted MAX-SAT algorithms to be used for clustering under the Crowley prior. In this paper we propose an extension in the context of spatial clustering.

3.1 Weighted MAX-SAT

A propositional atom, c_i, is created for each considered cluster c_i. Propositional atoms are binary variables with two values (TRUE and FALSE) and a partition is represented by setting all of its clusters to TRUE and all other clusters to FALSE. However, most truth-value assignments for the c_i do not correspond to a valid partition and, therefore, such assignments must be ruled out by constraints represented by logical clauses. A partition is a defined by clusters that do not overlap and where each data point must be included in some cluster.

A *propositional clause* is a *disjunction*: it states that at least one of a number of propositions is true and a formula which is a conjunction of clauses is said to be in conjunctive normal form (CNF). A weighted CNF formula is such that each clause has a positive weight attached. This weight should be interpreted as a cost which is incurred by an assignment if that assignment does not satisfy the clause. The total cost of any assignment is the sum of the costs of clauses that assignment fails to satisfy.

It is useful to allow some clauses to be hard clauses. Such clauses have infinite cost–they must be satisfied if at all possible. Other clauses are soft–it would be nice to satisfy them but an optimal assignment might break them. To encode the clustering problem we define one class of hard clauses and one class of soft clauses.

The first class of hard clauses states that each data point must be included in some cluster in the partition. Let $\{c_{y_1}, c_{y_2}, \ldots, c_{y_{i(y)}}\}$ be the set of all clusters containing

data point y. For each y a single clause of the form:

$$c_{y_1} \vee c_{y_2} \vee \cdots \vee c_{y_{i(y)}}$$

is created.

The second class of hard clauses rules out the inclusion of overlapping clusters we assert clauses of the form:

$$\overline{c_i} \vee \overline{c_j}$$

for all non-disjoint pairs of clusters c_i, c_j. (A bar over a formula represents negation.) Each such clause is logically equivalent to $\overline{c_i \wedge c_j}$: both clusters cannot be included in a partition.

The two classes of hard clauses above suffice to rule out non-partitions; it remains to ensure that each partition has the right score. This can be done by exploiting the decomposability of the partition score into cluster scores and using soft clauses to represent cluster scores. If S_i, the score for cluster c_i, is positive the following weighted clause is asserted:

$$S_i : c_i \tag{2}$$

If a cluster c_j has a negative score S_j then this weighted clause is asserted:

$$-S_j : c_j \tag{3}$$

which states a preference for c_j not to be included in the partition. Given an input composed of the clauses above the task of a weighted MAX-SAT solver is to find a truth assignment to the c_i which respects all hard clauses and maximizes the sum of the weights of satisfied soft clauses. Such an assignment will encode the highest scoring partition constructed from the given clusters.

See Liverani et al. (2010) for more details on the encoding of the necessary constraints and additional filters that can be included in the weighted MAX-SAT solver.

However, a full exploration of the partition space is not possible, as this space is huge. For example, there are 5.1×10^{13} ways to partition 20 elements and up to 1.04×10^6 cluster scores could be encoded into weighted MAX-SAT, considerably slowing down the solvers as n grows.

Liverani et al. (2010) suggest few alternatives to reduce the number of cluster scores, but in our context the geographical constraints imposed by spatial clustering can directly be encoded in the cluster score computation. In the next section we propose an algorithm that automatically includes such constraints in the clustering problem.

3.2 Computing Cluster Scores with Geographical Constraints

In spatial clustering a cluster is defined as a geographically bounded group of data points. Therefore, we propose to compute cluster scores only for such clusters, reducing the number of superfluous cluster scores that slow down the weighted MAX-SAT solvers.

A simple way to compute only cluster scores that satisfy the geographical constraints is to evaluate all the possible clusters containing a single observation and to iteratively augment the size of the most plausible clusters, thanks to the nice decomposability property of the score function. Note that we also need to include an upper bound for the size of each cluster, m, to avoid redundant clusters scores. We choose settings of the parameter $m \approx n/2$, as this guarantees speed in computation and it does not overly restrict the maximum size of any given cluster. However, this does not affect the results as we are unlikely to require, or expect, clusters of cardinality equal to, or greater than, $n/2$ for problems where clustering was deemed appropriate.

The algorithm to compute cluster scores is described below.

Step 1 Set the maximum cluster size, m. Set $i = 1$.
Step 2 Consider cluster $c_i = i$. Compute its cluster score.
Step 3 Consider all the data points that share a boundary with cluster c_i. Compute their cluster scores if they were to be merged with cluster c_i.
Step 4 Merge c_i with the data point that resulted in the highest increase in the partition score.
Step 5 Repeat Steps 3–4 until the cluster c_i reaches cardinality m.
Step 6 If $i = n$, go to Step 7. Else, set $i = i + 1$ and repeat Steps 2–6.
Step 7 Run a weighted MAX-SAT solver to identify the best partition.

4 Results on a Simulated Dataset

We include the results obtained on a small simulated dataset with the purpose of demonstrating the method proposed. We simulated four social indexes for the 20 regions of Tanzania.

All runs of weighted MAX-SAT were conducted using the C implementation that is available from the UBCSAT home page http://www.satlib.org/ubcsat. UBCSAT (Tompkins and Hoos 2005) is an implementation and experimentation environment for Stochastic Local Search (SLS) algorithms for SAT and MAX-SAT. We have used their implementation of WalkSat in this paper.

The weighted MAX-SAT solver retrieved the four generating clusters successfully. The left hand side of Fig. 1 shows the four geographical clusters obtained on the map of Tanzania, whilst the right hand side of Fig. 1 shows the values of the indicators for the observations that belong to the four clusters.

This example on a small simulated dataset showed that the methods described above can be implemented to produce sensible results.

5 Discussion

In this paper we proposed a modification to the algorithm by Liverani et al. (2010) for spatial clustering. We illustrated our method on a simulated dataset.

The method proposed presents many advantages. As mentioned above it is very flexible, so it can be modified to suit different types of data, such as social indexes

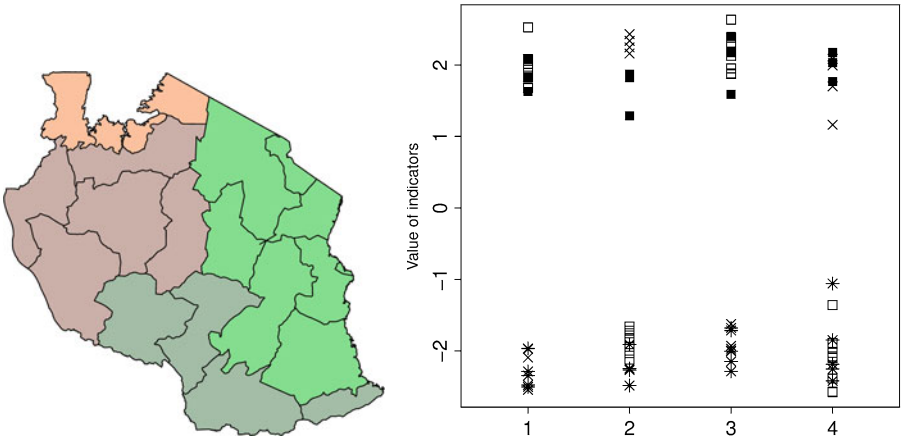


Fig. 1 On the left hand side, there is the map of the 20 regions of Tanzania. The different colors identify the regions which generated the data and were also retrieved by the algorithm. On the right hand side there are the values of the four indicators for the four clusters obtained by the algorithm. The four indicators are identified by different values of the x-axis, and the four clusters are identified by different symbols used in the plot

and measurements taken over time in a spatial context. Furthermore, the conjugacy of the proposed Bayesian model assures speed of the algorithm, and prior information can be included in the model.

Encoding the geographical constraints typical of spatial clustering for weighted MAX-SAT solvers is direct and sensible, by-passing the disadvantages of other widely used methods for partition space search, such as hierarchical clustering and stochastic search, that do not allow geographical constraints to be incorporated as naturally.

References

Crowley, E. M. (1997). Product partition models for normal means. *Journal of the American Statistical Association*, *92*(437), 192–198.

Cussens, J. (2008). Bayesian network learning by compiling to weighted MAX-SAT. In McAllester, D. A. and Myllymäki, P., editors, *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence*, pages 105–112, Helsinki, Finland. AUAI Press.

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, *41*, 578–588.

Gordon, A. D. (1996). A survey of constrained classification. *Computational Statistics & Data Analysis*, *21*(1), 17–29.

Heard, N. A., Holmes, C. C., & Stephens, D. A. (2006). A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves. *Journal of the American Statistical Association*, *101*(473), 18–29.

Johnston, R. J. (1976). *Classification in Geography: CLADAG 6*. Geo Abstracts.

Legendre, P. (1987). Constrained clustering. In P. Legendre & L. Legendre (Eds.), *Developments in numerical ecology* (pp. 289–307). Springer-Verlag, Berlin.

Liverani, S., Anderson, P. E., Edwards, K. D., Millar, A. J., & Smith, J. Q. (2009). Efficient utility-based clustering over high dimensional partition spaces. *Journal of Bayesian Analysis*, *4*(3), 539–572.

Liverani, S., Cussens, J. & Smith, J. Q. (2010). Searching a multivariate partition space using weighted MAX-SAT. F. Masulli, L. Peterson, and R. Tagliaferri (Eds.): Computational Intelligence Methods for Bioinformatics and Biostatistics, 6th International Meeting, CIBB 2009 Genova, Italy. Lecture Notes in Bioinformatics 6160, pp. 240–253, Springer, Berlin.

Ray, S., & Mallick, B. (2006). Functional clustering by Bayesian wavelet methods. *Journal of the Royal Statistical Society: Series B*, *68*(2), 305–332.

Smith, J. Q., Anderson, P. E., & Liverani, S . (2008). Separation measures and the geometry of bayes factor selection for classification. *Journal of the Royal Statistical Society: Series B*, *70*(5), 957–980.

Tompkins, D. A. D., & Hoos, H. H. (2005). UBCSAT: An implementation and experimentation environment for SLS algorithms for SAT and MAX-SAT. In H. H. Hoos & D. G. Mitchell (Eds.), *Theory and applications of satisfiability testing: Revised selected papers of the seventh international conference (SAT 2004, Vancouver, BC, Canada, May 10–13, 2004)* (pp. 306–320), Volume 3542 of *Lecture Notes in Computer Science*. Berlin, Germany: Springer Verlag.

Zhou, C., Wakefield, J. C., & Breeden, L. L. (2006). Bayesian analysis of cell-cycle gene expression data. In K. A. Do, P. Müller, & M. Vannucci (Eds.), Bayesian inference for gene expression and proteomics (pp. 177–200). Cambridge University Press, Cambridge.

Clustering Multiple Data Streams*

Antonio Balzanella, Yves Lechevallier, and Rosanna Verde

Abstract In recent years, data streams analysis has gained a lot of attention due to the growth of applicative fields generating huge amount of temporal data. In this paper we will focus on the clustering of multiple streams. We propose a new strategy which aims at grouping similar streams and, together, at computing summaries of the incoming data. This is performed by means of a divide and conquer approach where a continuously updated graph collects information on incoming data and an off-line partitioning algorithm provides the final clustering structure. An application on real data sets corroborates the effectiveness of the proposal.

1 Introduction

With the fast growing of capabilities in data acquisition and processing, a wide number of domains is generating huge amount of temporal data.

Some examples are financial and retail transactions, web data, network traffic, electricity consumptions, remote sensors data.

Traditional data mining methods fail at dealing with these data since they use computational intensive algorithms which require multiple scans of the data. Thus, whenever users need to get the answers to their mining and knowledge discovery queries in short times, such algorithms become ineffective.

A further issue of traditional algorithms is that data can be only processed if they are stored on some available media.

To deal with this new challenging task, proper approaches, usually referred as techniques for data streams analysis, are needed.

Among the knowledge extraction tools for data streams, clustering is widely used in exploratory analyses.

[1] This paper has been supported by COST Action IC0702 and by 'Magda una piattaforma ad agenti mobili per il Grid Computing' Chair: Prof. Beniamino Di Martino.

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_28,

Clustering in data stream framework is used to deal with two different challenges. The first one is related to analyzing a single data stream to discover a partitioning of the observations it is composed of. A second one is to process data streams generated by a set of sources (let's think about sensor networks) to discover a partitioning of the sources selves.

In this paper we will focus on the second one, which is usually referred as streams clustering.

Interesting proposals on this topic have been introduced in Beringer and Hullermeier (2006) and Dai et al. (2006). The first one is an extension to the data streams framework of the *k-means* algorithm performed on time series. Basically, the idea is to split parallel arriving streams into non overlapping windows and to process the data of each window performing, at first, a Discrete Fourier Transform to reduce the dimensionality of data, and then, the *k-means* algorithm on the coefficients of the transformation. On each window, the *k-means* is initialized using the centroids of the clusters of the partition obtained by the latest processed window.

The main drawback of this strategy is the inability to deal with evolving data streams. This is because the final data partition only depends from the data of the most recent window.

The second proposal is performed in two steps:

- an On-line procedure stores the coefficients of a suitable transformation (wavelet or linear regression) computed on chunks of the streams.
- an Off-line procedure is run on the on-line collected coefficients to get the final clustering structure.

Although this method is able to deal with evolving data streams, its main drawback is that the approach used for summarization is only based on storing compressed streams.

In this paper, we introduce a new strategy for clustering highly evolving data streams which provides the clustering structure over a specific temporal interval and a set of time located summaries of the data. Our proposal consists in two steps. The first one, which is run on-line, performs the clustering of incoming chunks of data to get local representative profiles and to update the adjacency matrix of an undirected graph which collects the similarities among the streams.

The second one, performs the final data partitioning by means of an off-line clustering algorithm that is run on the adjacency matrix.

2 A Graph Based Approach for Clustering Streaming Time Series

Let us note with $S = \{Y_1, \ldots, Y_i, \ldots, Y_n\}$ the n streams $Y_i = [(y_1, t_1), \ldots, (y_j, t_j), \ldots, (y_\infty, t_\infty)]$ made by real valued ordered observations on a discrete time grid $T = \{t_1, ..., t_j, ...t_\infty\} \in \Re$. A time window w_f with $f = 1, \ldots, \infty$ is an ordered

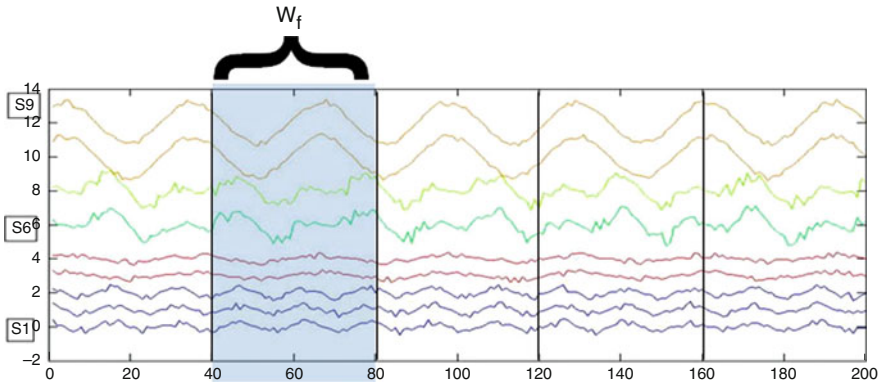


Fig. 1 Splitting of the incoming data into non overlapping windows

subset of T having size w_s. Each time window w_f frames a subset Y_i^w of Y_i called subsequence where $Y_i^w = \{y_j, \ldots, y_{j+w_s}\}$.

The objective is to find a partition P of S into C clusters such that each stream Y_i belongs to a cluster C_k with $k = 1, ..., C$ and $\bigcap_{k=1}^{C} C_k = \phi$. Streams are allocated to each cluster C_k with the aim to minimize the dissimilarity within each cluster and to maximize the dissimilarity between clusters.

In order to get the partition P, the incoming parallel streams are split into non overlapping windows of fixed size (Fig. 1).

On the subsequences Y_i^w of Y_i framed by each window w_f we run a Dynamic Clustering Algorithm (DCA) extended to complex data (Diday 1971, Carvalho et al. 2004).

The DCA looks both for a representation of the clusters (by means of a set of prototypes) and the best partition in K clusters, according to a criterion function based on a suitable dissimilarity measure.

The algorithm performs a step of representation of the clusters and a step of allocation of the subsequences to the clusters according to the minimum dissimilarity to the prototypes.

In our case, DCA provides a local partitioning $P_w = C_1^w \cup \ldots \cup C_\kappa^w \cup \ldots \cup C_K^w$ into K clusters of the subsequences framed by w_f and the associated set of prototypes $B^w = (b_1^w, \ldots, b_\kappa^w, \ldots, b_K^w)$ which summarize the behaviors of the streams in time localized windows.

Let $G = (V, E)$ be an undirected similarity graph where the vertex set $V = (v_1, \ldots, v_i, \ldots, v_n)$ corresponds to the indices i of the streams of S and the edges set E, carries non-negative values which stand for the weights $a_{i,l}$ of the linking between the streams Y_i and Y_l.

The graph G can be represented by means of a symmetric adjacency matrix $A = (a_{i,l})_{i,l=1,\ldots n}$ where $a_{i,l} = a_{l,i}$ and $a_{i,i} = 0$.

For each local partitions P_w we update the adjacency matrix A of the graph G, processing the outputs provided by the clustering algorithm at each window (Fig. 2).

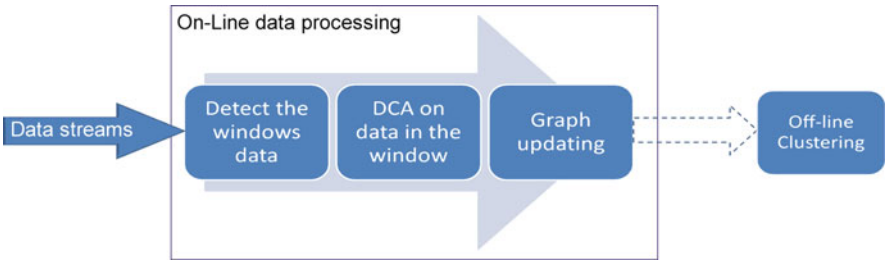


Fig. 2 Data processing schema

for each local cluster $C_\kappa^w \in P_w$ **do**
  Detect all the possible couples of subsequences Y_i^w, Y_l^w which are allocated to the cluster C_κ^w
  for each couple (i, l) **do**
    add 1 to the cells $a_{i,l}$ and $a_{l,i}$ of A
  end for
end for

This is performed by collecting the similarity values among the streams without computing all the pairwise proximities among the streams, as required in the data stream framework.

The main idea underlying this approach is to store in each cell $a_{i,l}$ the number of times each couple of streams is allocated to the same cluster of a local partition P_w.

This involves that the following procedure has to be run on each window:

For instance, let us assume to have five streams $(Y_1, Y_2, \ldots, Y_5)$ and a local partition $P_1 = (Y_1^w, Y_2^w)(Y_3^w, Y_4^w, Y_5^w)$, the graph updating consists in:

1. Adding 1 to the cells $a_{1,2}$ and $a_{2,1}$
2. Adding 1 to the cells $a_{3,4}$ and $a_{4,3}$
3. Adding 1 to the cells $a_{3,5}$ and $a_{5,3}$
4. Adding 1 to the cells $a_{4,5}$ and $a_{5,4}$

We can look at A as a proximity matrix where the cells record the pairwise similarity among streams.

For each couple of subsequences Y_i^w, Y_l^w in the same cluster, the updating of the cells $a_{i,l}$ and $a_{l,i}$ of A with the value 1, corresponds to assign the maximum value of similarity to such couple of subsequences.

When this procedure is performed on a wide number of windows, it is a graduation of the consensus between couple of streams computed incrementally. The higher is the consensus, the more similar will be the streams.

According to the proposed procedure, if two subsequences belong to two different clusters of a local partition, their similarity is always considered minimal, with the value 0 set in the corresponding cells of A.

In order to deal with this issue, we improve the previous graph updating approach, by graduating the similarities between the couples of streams belonging to different clusters instead to consider these maximally dissimilar.

To reach this aim, we define the *cluster radius* and the *membership function*.

Definition 1. Let $d \in \Re$ be a distance function between two subsequences. The radius rad_κ^w of the cluster C_κ^w is $max(d(Y_i^w, b_\kappa^w))$ for each $Y_i^w \in C_\kappa^w$.

Starting from the cluster radius rad_κ^w, the *Membership function* of a subsequence to a cluster C_κ^w is given by:

$$mf_{l,\kappa} = \frac{rad_\kappa^w}{d(Y_l^w, b_\kappa^w)} \quad \forall Y_l^w \notin C_\kappa^w \tag{1}$$

$mf_{l,\kappa}$ ranges from 0 to 1 since it is the ratio between the radius of the cluster C_k and the distance of a subsequence Y_l^w not belonging to the cluster from the prototype $b_\cup^w$ of C_k^w.

We update the cells of A using the value of the *Membership function* computed on each couple subsequences belonging to different clusters.

According to such adjustments, the whole updating process becomes the following:

Algorithm 1 Graph updating strategy

```
for all Y_i^w   i ∈ n do
  Detect the index κ of the cluster which Y_i^w belongs to
  for all Y_l^w ∈ C_κ^w do
    Add the value 1 to the graph edges a_{i,l} and a_{l,i}
  end for
  for all γ ≠ κ   γ = 1, ..., K do
    Compute mf_{i,γ}
    Detect all the subsequences Y_l^w ∈ C_γ^w
    Add the value mf_{i,γ} to the cells a_{i,l} and a_{l,i} of A
  end for
end for
```

The proposed procedure, when repeated on each window, allows to get the pairwise similarities among streams in an incremental way. This is obtained without computing the similarity between each couple of streams because only the distances computed in the allocation step of the last iteration of the DCA algorithm are used.

3 Off-line Graph Partitioning by Spectral Clustering

In order to get a global partition P into C clusters, of the set of streams S analyzed in the sequence of windows w_f ($f = 1, \ldots, F$), we have to run a proper clustering algorithm on the proximity matrix A.

The choice of the F value of processed windows, is performed according to a user clustering demand or to a time schedule which breaks the updating of A at prefixed time points such to get the clustering results over several time horizons.

Usual techniques for clustering proximity graphs are based on spectral clustering procedures (Luxburg 2007).

The problem of finding a partition of S where the elements of each group are similar among them while elements in different groups are dissimilar can be formulated in terms of similarity graphs. We want to find a partition of the graph such that the edges between different groups have a very low weight (similarity) and the edges within a group have high weight (similarity).

Note that according to Bavaud (2006), similar results can be obtained using a non metric multidimensional scaling on the adjacency matrix A.

Definition 2. Let $v_i \in V$ a vertex of the graph. The degree of v_i is $d_i = \sum_{l=1}^{n} a_{i,l}$. The degree matrix D is defined as the diagonal matrix with the degrees $d_1, \ldots, d_n$ on the diagonal.

Starting from the degree matrix D it is possible to define the Laplacian Matrix and the Normalize Laplacian Matrix:

Definition 3. Let $L = D - A$ be the unnormalized Laplacian Matrix. The normalized Laplacian Matrix L_{norm} is defined as:

$$L_{norm} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} \tag{2}$$

The algorithm schema is the following:

Algorithm 2 Spectral clustering algorithm

Compute L and L_{norm}
Compute the first C eigenvalues of L_{norm}
Let $\tilde{A} \in \Re^{nxC}$ the matrix containing the C eigenvectors associated to the first eigenvalues of L_{norm}
Get the partition P of S into C clusters using the k-means algorithm on the multidimensional points defined by the rows of $\tilde{A}$

4 Main Results

To evaluate the performance of the proposed strategy we have compared the clustering performance of the on-line clustering strategy with the k-means algorithm on stocked data, using highly evolving datasets.

Two datasets have been used in the evaluation process:

The first one is made by 76 highly evolving time series, downloaded from Yahoo finance, which represent the daily closing price of random chosen stocks. Each time series is made by 4,000 observation.

The second one is made by 179 highly evolving time series which collect daily electricity supply at several locations in Australia. Each time series is made by 3,288 recordings.

We have considered some of the most common indexes to assess the effectiveness of the proposal (See Maulik and Bandyopadhyay 2002). The Rand index (RI) and the Adjusted Rand index (ARI) are used as external validity indexes to evaluate the degree of consensus between the partition obtained by our proposal and the partition obtained using the k-means. Moreover, the Calinski-Harabasz Index(CH), the Davies-Bouldin(DB) Index and the Silhouette Width Criterion(SW), are used as internal validity indexes to evaluate the compactness of clusters and their separation.

In order to perform the testing, we need to set the following input parameters for the proposed procedure:

- number of clusters K of each local partition P_w
- the final number of cluster C to get the partition of S
- the size w_f of each temporal window

For the k-means we only need to set the number of clusters C.

The Euclidean distance is used as dissimilarity function in both the procedures.

According to this choice, DCA algorithm on the windows data, is a classical *k-means* where the prototypes are the average of the data in a cluster.

The parameter C has been set, for the first and second datasets, running the *k-means* algorithm using $C = 2, \ldots, 8$. For each value of C we have computed the total within deviance.

We have chosen $C = 4$ for the first dataset and $C = 3$ for the second dataset, since these are the values which provide the highest improvement of the clusters homogeneity.

By evaluating, through the mentioned indexes, the partitioning quality for several values of w_s we can state that the choice of the windows size does not impact on the clusters homogeneity. As a consequence, the choice of the value of such parameter, can be performed according to the kind of required summarization. For example, if we need to detect a set of prototypes for each week of data, we choose a value of the window size which frames the observations in a week.

In our tests, we have used windows made by 30 observations for the first two datasets and 50 for the third one.

The third required input parameter K does not strongly impact on the clustering quality. We have tested this by evaluating the behavior of the Calinski-Harabasz Index and of the Davies-Bouldin Index according to $k = 2, \ldots, k = 10$.

In Table 1 we show the main results for the evaluated indexes.

Looking at the values of the internal validity indexes, computed for our proposal and for the k-means on stocked data, it emerges that the homogeneity of the clusters and their separation, is quite similar.

Table 1 External and internal validity indices

Dataset	On-line clustering					K-means clustering		
	DB	CH	SW	RI	ARI	DB	CH	SW
Power supply	2.049	26.013	0.223	0.92	0.83	2.172	26.504	0.229
Financial data	1.793	14.39	0.270	0.88	0.80	1.754	15.594	0.321

Moreover, the value of the Rand Index and of the Adjusted Rand Index, highlights the strength of the consensus between the obtained partitions.

5 Conclusions and Perspectives

In this paper we have introduced a new strategy for clustering highly evolving data streams based on processing the incoming data in incremental way without requiring their storage. This strategy favorably compares to the standard k-means performed on stocked data as it is shown on the test datasets using several standard validity indexes. Further developments will be to introduce a strategy for monitoring the evolution of the clustering structure over time.

References

Bavaud, F. (2006). Spectral clustering and multidimensional scaling: A unified view. In V. Batagelj, H.-H. Bock, A. Ferligoj, & A. Ziberna (Eds.), *Data science and classification* (pp. 131–139). New York: Springer.

Beringer, J., & Hullermeier, E. (2006). Online clustering of parallel data streams. *Data and Knowledge Engineering*, *58*(2), 180–204.

Bi-Ru Dai, Jen-Wei Huang, Mi-Yen Yeh, & Ming-Syan Chen (2006). Adaptive Clustering for Multiple Evolving Streams. *IEEE Transactions on Knowledge and Data Engineering*, *18*(9), 1166–1180.

De Carvalho, F., Lechevallier, Y., & Verde, R. (2004). Clustering methods in symbolic data analysis. In D. Banks, L. House, F. R. McMorris, P. Arabie, & E. Gaul (Eds.), *Classification, clustering, and data mining applications. Studies in classification, data analysis, and knowledge organization* (pp. 299–317). Berlin: Springer.

Diday, E. (1971). La methode des Nuees dynamiques. *Revue de Statistique Appliquee*, *19*(2), 19–34.

Maulik, U., & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(12), 1650–1654.

von Luxburg, U. (2007, December). A tutorial on spectral clustering. *Statistics and Computing*, *17*(4), 395–416.

Notes on the Robustness of Regression Trees Against Skewed and Contaminated Errors

Giuliano Galimberti, Marilena Pillati, and Gabriele Soffritti

Abstract Regression trees represent one of the most popular tools in predictive data mining applications. However, previous studies have shown that their performances are not completely satisfactory when the dependent variable is highly skewed, and severely degrade in the presence of heavy-tailed error distributions, especially for grossly mis-measured values of the dependent variable. In this paper the lack of robustness of some classical regression trees is investigated by addressing the issue of highly-skewed and contaminated error distributions. In particular, the performances of some non robust regression trees are evaluated through a Monte Carlo experiment and compared to those of some trees, based on M-estimators, recently proposed in order to robustify this kind of methods. In conclusion, the results obtained from the analysis of a real dataset are presented.

1 Introduction

Tree-based regression represents a simple and widely-used alternative to parametric regression. It is very popular in data mining applications (see for example Azzalini and Scarpa 2004; Hastie et al. 2009), in which datasets are often very large and may have different kinds of variables and many missing values. Binary regression trees (Breiman et al. 1984) approximate an unknown regression function by a step function defined on the covariate space. This goal is obtained by means of a procedure that recursively bisects the data into a number of disjoint subgroups. The procedure requires the definition of a splitting method that drives the growing phase of the tree, a pruning procedure and constant values for predicting the response variable Y within each terminal node.

The most widely-employed tree-based procedure relies on a least squares (OLS) criterion, in which the data are split by minimizing the within-node sum of the squared errors. Trees obtained using this criterion clearly concentrate on modelling the relationship between the response and the covariates at the centre (i.e., the conditional mean) of the response distribution. They also provide piecewise constant estimates of the regression function. Despite their wide use, properties of OLS trees

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_29,

have not been deeply explored in the literature. For example, even though OLS trees do not make any assumption on the error term distribution, the OLS split criterion corresponds to a maximum likelihood split criterion in which normally distributed error terms are assumed (Su et al. 2004). For this reason, OLS trees result to be optimal under conditions which are not always true, especially in predictive data mining applications. For example, a few unusually low or high values of Y may have a large influence on the value of the residual sum of squares (Breiman et al. 1984, Galimberti et al. 2007).

The lack of robustness of regression trees may be addressed following two different approaches. The first, referred to as the identification/rejection approach, consists of identifying outliers and influential observations and removing them from the sample before performing the data analysis. The second, referred to as robust regression, tries to devise suitable modifications of the standard methodologies so that the results obtained from the analysis of the entire sample are not so affected by the presence of outliers and influential observations. Following the first approach, a solution can be obtained, for example, by including the forward search strategy (Atkinson and Riani 2000) in tree-based regression. This idea has been explored in Mascia et al. (2005). A method for identifying outliers in the context of classification trees which is based on the identification/rejection approach has been proposed in John (1995). Within the second approach, tree-structured regression methods were combined with some fundamental ideas in piecewise polynomial quantile regression (Chaudhuri and Loh 2002). The resulting method is able to provide insight into the centre as well as the lower and upper tails of the conditional distribution of Y, and thus can be useful whenever the covariates have different effects on different parts of the conditional distribution of Y. This method includes the piecewise constant median regression trees constructed using least absolute deviations (LAD trees), firstly introduced in Breiman et al. (1984) as a robust alternative to OLS trees.

The use of regression trees to predict skewed or heavy-tailed dependent variables has been considered already in a study whose aim was to evaluate the effects of social and economic factors on household income (Costa et al. 2006). In that study OLS trees tended to build partitions of the set of households in which a terminal node was composed of most of the units with a low income, while all the other terminal nodes comprised only a few households with large income values. In order to avoid this drawback an alternative solution within the robust regression approach was proposed, which is based on the Gini index and is suitable for analysing transferable dependent variables. It allowed for a better description of the factors which characterize low income households.

In order to overcome the lack of robustness of OLS trees against the presence of outlying values in the dependent variable, some robust regression trees based on M-estimation methodology (Maronna et al. 2006) were recently proposed and compared to OLS and LAD trees (Galimberti et al. 2007). In particular, M-estimators based on Huber's and Tukey's methods were employed. A Monte Carlo study was also performed in which outlying values in the dependent variable were generated using error terms distributed as a mixture of two normals with the same means but

with different variances. In that study the OLS criterion led to over-simplified trees with large biases. Moreover, LAD trees resulted to be actually more robust than OLS trees but not as much as Huber's and Tukey's trees (for further details see Galimberti et al. 2007).

In this paper the issue of analyzing skewed dependent variables with regression trees is further investigated. In particular, some studies based on simulated and real datasets are performed whose aim is to compare the performances of OLS, LAD, Huber's and Tukey's regression trees with highly skewed and contaminated datasets. Section 2 recalls the use of M-estimation methodology in regression tree building. Section 3 shows the results of a Monte Carlo experiment in which data are generated with outlying values in the dependent variable using schemes different from the one examined in Galimberti et al. (2007), namely contaminated, non-contaminated normal and log-normal error terms. The results from the analysis of a real data set concerning durations of phone calls (Azzalini and Scarpa 2004) as a typical predictive data mining application are illustrated in Sect. 4. Finally, Sect. 5 contains some concluding remarks.

2 Regression Trees Based on M-estimators

OLS and LAD tree-based regression procedures were recently generalised using the M-estimation methodology (Galimberti et al. 2007). This generalization was obtained by constructing trees whose nodes are iteratively split so as to maximize the decrease in the following objective function:

$$\sum_{t \in T} \sum_{i \in t} \rho\left(y_i - \theta_t\right), \tag{1}$$

where T denotes the set of terminal nodes at a given step of the splitting process, y_i is the response value measured on the ith statistical unit belonging to node t, and θ_t is the M-estimate of the location parameter within node t obtained by minimizing a given function $\rho(\cdot)$. The OLS and LAD criteria are special cases of the objective function (1), for $\rho(y_i - \theta_t) = (y_i - \theta_t)^2$ and $\rho(y_i - \theta_t) = |y_i - \theta_t|$, respectively. Two other popular choices for $\rho(\cdot)$ are Huber's and Tukey's functions, which lead to the so-called Huber's and Tukey's estimators. Both types of functions depend on a tuning constant k which is used to distinguish small from large absolute values of $y_i - \theta_t$. Huber's functions are defined as

$$\rho_{HU}(y_i - \theta_t)_k = \begin{cases} (y_i - \theta_t)^2 & \text{if } |y_i - \theta_t| \leq k, \\ 2k|y_i - \theta_t| - k^2 & \text{if } |y_i - \theta_t| > k. \end{cases} \tag{2}$$

Tukey's functions are defined as

$$\rho_{TU}(y_i - \theta_t)_k = \begin{cases} 1 - \left\{1 - [(y_i - \theta_t)/k]^2\right\}^3 & \text{if } |y_i - \theta_t| \leq k, \\ 1 & \text{if } |y_i - \theta_t| > k. \end{cases} \tag{3}$$

It is interesting to note that each M-estimator can be expressed as a weighted mean of the response values. In particular, OLS estimators equally weight all data

points, while Huber's and Tukey's estimators provide adaptive weighting functions that assign smaller weights to observations with large residuals: the weighting function corresponding to the Huber's estimator declines when $|y_i - \theta_t| > k$; the weights for the Tukey's estimator decline as soon as $|y_i - \theta_t|$ departs from 0, and are set equal to 0 for $|y_i - \theta_t| > k$ (for further details see Galimberti et al. 2007; Maronna et al. 2006).

As far as the choice of k is concerned, small values produce high resistance to outliers, but at the expense of low efficiency when the errors are i.i.d. normal variables with zero mean and variance σ^2. Thus, the constant k is generally selected to give reasonably high efficiency in the normal case. In particular, setting $k = 1.345\hat{\sigma}$ in the Huber's function and $k = 4.685\hat{\sigma}$ in the Tukey's one produces 95% asymptotic efficiency when the errors are normal, while still offering protection against outliers (Maronna et al. 2006). The quantity $\hat{\sigma}$ is a previously computed dispersion estimate of the error term. A robust way to obtain such an estimate may be to compute the median of the absolute values of the differences from the median (also referred to as the median absolute deviation), divided by 0.675. This is an approximately unbiased estimator of σ if n is large and the error distribution is normal (Maronna et al. 2006). The results described in the following Sections were obtained by applying this estimator to the residuals computed from the fitting of the LAD tree to each sample, and using the above mentioned values for k. Finally, it is worth mentioning that the pruning procedure for each robust strategy has been based on a cost-complexity measure defined according to the same criterion used in the growing phase.

3 A Simulation Experiment with Skewed and Contaminated Errors

A Monte Carlo experiment has been performed using the statistical software system `R` (R Development Core Team 2010) in order to investigate the behavior of OLS, LAD, Huber's and Tukey's regression trees in some situations in which the assumption of Gaussian error terms is violated. The recursive procedures whose split criteria are based on Huber's and Tukey's regression trees were implemented in a `R` code by suitably modifying the `R` package `rpart`. In particular, this experiment focuses on skewed and contaminated error term distributions. Data were generated according to the following model:

$$Y = f(X_1, X_2) + \varepsilon + c \cdot \xi \tag{4}$$

where $f(\cdot)$ is the step function with nine steps depicted in Fig. 1; X_1 and X_2 are i.i.d. random variables such that $X_h \sim Unif(-1, 1)$, for $h = 1, 2$; ε is a standardized random error term; ξ is a binary random contamination term with $\xi \sim Ber(\delta)$ (independent from X_1, X_2 and ε), and $c > 0$ is a fixed constant that controls the contamination intensity. It should be noted that the step function and the predictors

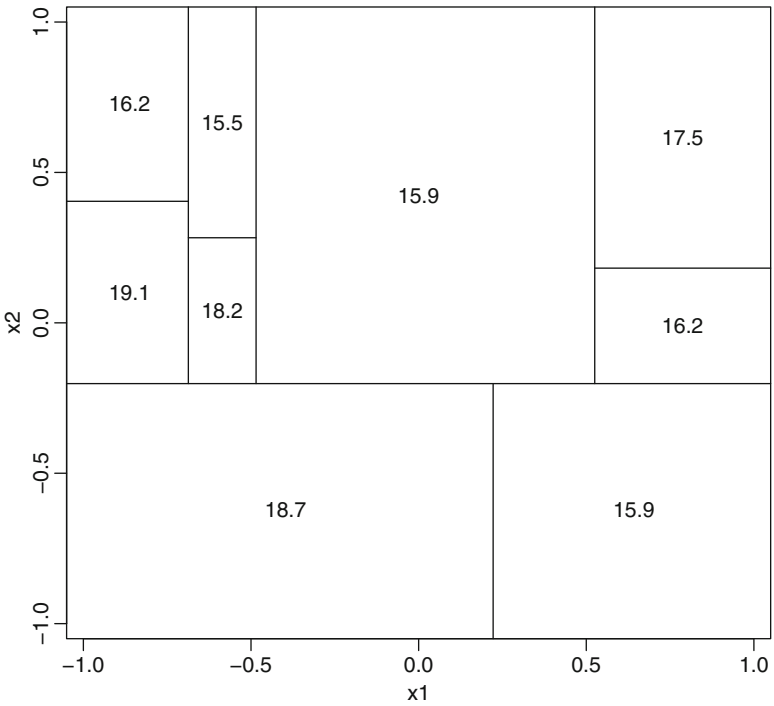


Fig. 1 Values of $f(X_1, X_2)$ and corresponding partition of the predictor space used in the simulation experiment

used in this experiment are exactly the same already considered in Galimberti et al. (2007), while the remaining terms of model (4) are different.

The factors considered in this study were: the shape of the error term distribution (normal or log-normal with skewness 1.5 and 3); the value of the contamination probability ($\delta = 0,\ 0.01,\ 0.05$); the value of the contamination intensity ($c = \min f(X_1, X_2),\ \max f(X_1, X_2)$), and the number of sample units ($n = 500,\ 1000,\ 2000$). For each combination of the levels of these factors, 50 learning samples and 50 test samples of n units were generated from model (4). Then, each couple of learning and test samples were used to construct and prune a regression tree through the four strategies illustrated in Sect. 2. Each pruned tree was evaluated with a special emphasis on the following two aspects: the number of terminal nodes (tree size) and the mean squared error. This latter measure was estimated as the mean of the squared differences between the values of the true step function $f(.)$ used to generate the data and the values of the estimated function $\hat{f}(.)$ given by the tree, evaluated on a uniform grid of 100×100 values for $(X_1,\ X_2)$.

Table 1 shows the estimated mean squared errors computed over 50 trees for $c = \min f(X_1, X_2)$, $n = 1000$ and for each combination of the other factors. For the same trees, Table 2 shows their mean tree sizes. As expected, when $\delta = 0$ and the

Table 1 Estimated mean squared errors of four regression tree-based methods (over 50 trees) (standard deviations in brackets) for different levels of δ and different error term distributions ($n = 1000$)

δ	Method	Normal	Log-normal	
			Skewness = 1.5	Skewness = 3
0.00	OLS	0.095 (0.039)	0.095 (0.042)	0.101 (0.044)
	LAD	0.137 (0.055)	0.160 (0.054)	0.175 (0.051)
	HUBER	0.117 (0.050)	0.133 (0.045)	0.142 (0.045)
	TUKEY	0.122 (0.045)	0.128 (0.043)	0.155 (0.060)
0.01	OLS	0.265 (0.107)	0.271 (0.103)	0.264 (0.109)
	LAD	0.134 (0.052)	0.153 (0.054)	0.172 (0.053)
	HUBER	0.119 (0.049)	0.129 (0.043)	0.141 (0.044)
	TUKEY	0.122 (0.051)	0.129 (0.042)	0.152 (0.057)
0.05	OLS	1.341 (0.385)	1.316 (0.380)	1.343 (0.375)
	LAD	0.181 (0.160)	0.161 (0.072)	0.172 (0.071)
	HUBER	0.146 (0.050)	0.140 (0.075)	0.144 (0.062)
	TUKEY	0.120 (0.045)	0.127 (0.048)	0.152 (0.067)

Table 2 Mean tree sizes of four regression tree-based methods (over 50 trees) (standard deviations in brackets) for different levels of δ and different error term distributions ($n = 1000$)

δ	Method	Normal	Log-normal	
			Skewness = 1.5	Skewness = 3
0.00	OLS	10.12 (1.66)	10.54 (2.56)	10.10 (1.68)
	LAD	10.34 (2.62)	10.50 (2.35)	11.06 (2.39)
	HUBER	10.14 (2.12)	10.60 (2.40)	11.04 (2.43)
	TUKEY	10.00 (2.38)	10.90 (3.53)	11.70 (2.69)
0.01	OLS	8.64 (2.44)	8.78 (2.76)	8.24 (2.09)
	LAD	10.12 (2.62)	10.36 (2.61)	10.74 (2.33)
	HUBER	10.10 (2.34)	10.56 (2.37)	11.14 (2.59)
	TUKEY	10.34 (3.07)	10.80 (3.90)	11.74 (2.65)
0.05	OLS	5.54 (2.04)	5.46 (1.98)	5.42 (2.11)
	LAD	9.78 (2.45)	10.00 (2.63)	10.18 (2.46)
	HUBER	9.76 (1.96)	10.08 (2.44)	10.04 (2.17)
	TUKEY	10.48 (2.62)	10.36 (2.74)	11.28 (2.86)

error term has a normal distribution, the lowest mean squared error is obtained with OLS regression trees. The introduction of skewness in the error term distribution seems to have little impact on the overall prediction performances of OLS trees, while it has greater effects on the performances of the other tree methods. However, even for small values of δ, there is a dramatic increase in the mean squared errors of OLS trees, for each distributional shape of the error terms. On the contrary, the performances of the other three methods result to be unaffected by the presence of contaminated data. Furthermore, Tukey's and Huber's trees perform better than LAD trees. As far as the differences in the tree sizes are concerned, in the presence of contaminated data using the OLS criterion leads to a reduction in tree sizes, which

may be related to the increase in the mean squared errors. Similar conclusions hold true when $c = \max f(X_1, X_2)$, $n = 500$ and $n = 2000$.

4 Predicting Durations of Phone Calls

The procedures for constructing regression trees described in Sect. 2 were also applied to a dataset containing information about $n = 30619$ customers of a mobile telephone company (Azzalini and Scarpa 2004). For each customer, the values of the following variables are available for each of ten consecutive months: the number of made and received phone calls, their total duration, the total cost of the made phone calls, the number of SMS, and the number of phone calls made to the customer care service. The purpose of this analysis is to predict the total duration of the made phone calls during the 10th month period based on the variables concerning the previous 9 months. The distribution of the dependent variable in the sample is highly skewed. Furthermore, many customers have a value equal to 0 for this dependent variable.

The dataset was split into two subsets: 15310 randomly selected customers were used as a training set to build the regression trees, and the remaining customers were employed as a test set to choose the optimal size of each tree. Table 3 summarizes some features of the pruned trees. A first difference among the considered trees concerns their sizes, ranging from 8 (OLS) to 54 (LAD). This result seems consistent with the results of the above mentioned studies on data with a heavy-tailed distribution for the dependent variable (Costa et al. 2006; Galimberti et al. 2007), and also with the results described in Sect. 3. Furthermore, OLS and Tukey's trees focus on opposite parts of the Y range. In the OLS tree 93.5% of the customers are assigned to the same terminal node, whose predicted value of Y is 843.1 (the lowest predicted value from this tree). The remaining 6.5% of the customers are split into seven terminal nodes, whose predicted values range from 5762.1 to 116733.6 (see the second row of Table 3). Thus, the OLS tree mainly focuses on the customers whose value of Y is high. On the contrary, Tukey's tree assigns 83.0% of the customers to seven terminal nodes, whose predicted values are lower than 811.7. For the remaining 19 terminal nodes the predicted values range from 1189.1 to 21668.4. Thus, the Tukey's tree

Table 3 Descriptive statistics of Y and $\hat{Y}$ according to the four pruned regression trees obtained from the analysis of the phone calls dataset (test set)

Y	Min	1st quartile	2nd quartile	3rd quartile	Max	Tree size
	0.0	0.0	165.0	948.0	168407.0	
OLS	843.1	843.1	843.1	843.1	116733.6	8
LAD	0.0	0.0	78.0	897.0	105659.0	54
HUBER	142.8	142.8	142.8	1180.4	105659.0	27
TUKEY	53.8	53.8	190.9	811.3	21668.3	26

predicts the lowest total durations of the phone calls made during the 10th months in a very precise way. As far as LAD and Huber's trees are concerned, they seem to be less focused on specific parts of the Y range. With Huber's tree 54.9% of the customers are assigned to the terminal node with the lowest predicted value, equal to 142.8. A distinctive feature of LAD tree is that it assigns 0 as predicted value to 35.1% of the customers.

5 Concluding Remarks

The examples illustrated in this paper show that, when the distribution of the dependent variable is highly-skewed and/or data are contaminated, the choice of the split criterion used to build the tree plays a fundamental role. In particular, OLS trees seem to be little affected by the skewness of the error distribution, but their performances severely degrade in the presence of contaminated data. On the contrary, LAD, Huber's and Tukey's trees are robust only against contaminated data but not against skewness. Thus, the choice of the split criterion requires careful attention. Since it may be difficult to establish a priori the best objective function for the problem at hand, a possible criterion could be based on the comparison of the tree sizes. As the simulation study shows, the four techniques considered in this paper lead to trees with similar sizes when data are not contaminated; in this case, the trees with the best performances are the OLS ones. Remarkable differences in the tree sizes can be related to the presence of contaminated data; in this situation, Huber's or Tukey's trees should be preferred. An alternative strategy could be the definition of locally optimal objective functions through an a posteriori analysis of regression trees obtained from different split criteria. This perspective has not been explored in the literature yet, and may represent a promising direction of future research.

References

Atkinson, A., & Riani, M. (2000). *Robust diagnostic regression analysis*. New York: Springer.

Azzalini, A., & Scarpa, B. (2004). *Analisi dei dati e data mining*. Milano: Springer.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Belmont: Wadsworth.

Chaudhuri, P., & Loh, W-Y. (2002). Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, *8*, 561–576.

Costa, M., Galimberti, G., & Montanari, A. (2006). Binary segmentation methods based on Gini index: A new approach to the multidimensional analysis of income inequalities. *Statistica & Applicazioni*, *IV*, 123–141.

Galimberti, G., Pillati, M., & Soffritti, G. (2007). Robust regression trees based on M-estimators. *Statistica*, *LXVII*, 173–190.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.

John, G. H. (1995). Robust decision trees: Removing outliers from databases. In U. M. Fayyad & R. Uthurusamy (Eds.), *Proceedings of the first international conference on knowledge discovery and data mining (KDD-95)* (pp. 174–179). Montreal: AAAI Press.

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics. Theory and methods*. New York: Wiley.

Mascia, P., Miele, R., & Mola, F. (2005). Outlier detection in regression trees via forward search. In S. Zani & A. Cerioli (Eds.), *Proceedings of the meeting of the classification and data analysis group of the Italian statistical society* (pp. 429–432). Parma: Monte Università Editore.

R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. `URL http://www.R-project.org`.

Su, X., Wang, M., & Fan, J. (2004). Maximum likelihood regression trees. *Journal of Computational and Graphical Statistics*, *13*, 586–598.

A Note on Model Selection in STIMA

Claudio Conversano

Abstract Simultaneous Threshold Interaction Modeling Algorithm (STIMA) has been recently introduced in the framework of statistical modeling as a tool enabling to automatically select interactions in a Generalized Linear Model (GLM) through the estimation of a suitable defined tree structure called 'trunk'. STIMA integrates GLM with a classification tree algorithm or a regression tree one, depending on the nature of the response variable (nominal or numeric). Accordingly, it can be based on the Classification Trunk Approach (CTA) or on the Regression Trunk Approach (RTA). In both cases, interaction terms are expressed as 'threshold interactions' instead of traditional cross-products. Compared with standard tree-based algorithms, STIMA is based on a different splitting criterion as well as on the possibility to 'force' the first split of the trunk by manually selecting the first splitting predictor. This paper focuses on model selection in STIMA and it introduces an alternative model selection procedure based on a measure which evaluates the trade-off between goodness of fit and accuracy. Its performance is compared with the one deriving from the current implementation of STIMA by analyzing two real datasets.

1 Introduction

Generalized Linear Model (GLM) is a flexible generalization of ordinary least squares regression. The theory underlying GLM was introduced in Nelder and Wedderburn (1972). It is based on the hypothesis that the response variable follows a member of the single-parameter exponential family of probability distributions. Depending on the nature of the data, this distribution relates to the gaussian, binomial, Poisson, gamma, inverse gaussian, geometric and negative distribution. In most cases the response is modeled as, or can be expressed as, a linear combination of the predictors.

A difficult task in GLM is to find interactions between two or more predictors. Their presence, and their consequent inclusion in the estimated model, may be a key to correct interpretation of data. From a statistical perspective, interaction between two or more predictors occurs if their separate effects do not combine additively

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_30,

(Gonzalez and Cox 2007) or, equivalently, when over and above any additive combination of their separate effects, they have a joint effect (Cohen et al. 2003). As a result, interaction is a particular kind of non-additivity which is usually modeled as a cross-product between two or more predictors.

Simultaneous Threshold Interaction Modeling Algorithm (STIMA) (Conversano and Dusseldorp 2010) has been recently introduced to deal with interaction modeling in GLM. It combines GLM and CART-like recursive partitioning algorithm (Breiman et al. 1984): GLM is used to estimate the main effect of the model, whereas recursive partitioning is used to estimate simultaneously the interaction effect. Depending on the nature (nominal or numeric) of the response variable, a data-driven method named Classification Trunk Approach (CTA) (Dusseldorp and Meulman 2004) or Regression Trunk Approach (RTA) (Dusseldorp et al. 2010) is used to identify interaction terms and the resulting model is named classification (or regression) trunk.

This paper deals with model selection in STIMA, and it introduces an alternative approach based on a measure which evaluates the trade-off between goodness of fit and accuracy. In the following, Sect. 2 describes the trunk model and summarizes the main features of the STIMA algorithm. Section 3 focuses on model selection, whereas Sect. 4 evaluates the performance of STIMA with the 'new' model selection criteria. Concluding remarks are reported in Sect. 5.

2 STIMA

2.1 The Trunk Model

The model estimated by STIMA is equivalent, in its formulation, to a standard GLM. Dealing with J (numeric or nominal) predictors x_j $(j = 1, \ldots, J)$, STIMA assumes the response y has an exponential family density $\rho_y(y; \theta; \phi)$ with a natural parameter θ and a scale parameter ϕ; the mean $\mu = E(y|x_1, \ldots, x_J)$ is linked to the x_js via:

$$g(\mu)_L = \underbrace{\beta_{0L} + \sum_{j=1+d}^{J} \beta_{jL} x_j}_{\textit{main effect}} + \underbrace{\sum_{m=1}^{M_L - 1} \beta_{j+mL}\, I\left\{(x_j \leq s_j, \ldots, x_1 < s_1) \in R_{mL}\right\}}_{\textit{interaction effect}} \tag{1}$$

The first term relates to the main effect model estimated via GLM and the second to the interaction effect estimated using recursive partitioning. Main idea is to fit a classification tree or a regression tree over and above the linear main effect of predictors to identify interaction effects. Since the algorithm generates reduced-size trees, the resulting tree and the overall estimation method are named classification/regression trunk and Classification/Regression Trunk Approach (CTA or RTA) respectively.

Notation used in (1) is consistent with that used in (Hastie et al. 2001): L is the number of splits of the trunk and M_L is the total number of its terminal nodes; $\hat{\beta}_{j+mL}$ is the response outcome assigned to observations falling into the terminal node R_m ($i = 1, \ldots, M$); d equals zero if the first splitting predictor is numeric or one if it is nominal. The indicator function $I(\cdot)$ assigns observations to one of the terminal nodes based on the splitting values s_j of the splitting predictors x_j ($j = 1, \ldots, J$). The total number of indicator variables $I(\cdot)$ included in the model equals $M_L - 1$, since one of the terminal nodes of the trunk serves as reference group. As a result, (1) points out M_{L-1} threshold interaction terms that are automatically identified by the trunk and the final model includes $J + M_L$ estimated parameters. If the response is numerical, the link function $g(\cdot)$ is the identity function and (1) reduces to the regression trunk model introduced in Dusseldorp et al. 2010. Whereas, a nominal response leads to a GLM with threshold interaction terms. The specification and the features of such a model for the binary response case are discussed in Conversano and Dusseldorp (2010).

2.2 Algorithm

STIMA implements the estimation algorithm for CTA or RTA and consists of a trunk-growing step and a trunk-pruning step.

In the ℓ-th iteration of the trunk-growing process ($\ell = 1, \ldots, L$), the ℓ-th interaction term entering the model is the one maximizing the effect size $f^{(\ell)}$, i.e., the relative decrease in the residual deviance when passing from the model with $\ell - 1$ terms to the one with ℓ terms. In practice, for each possible combination of splitting variable x_j, split point s_j and splitting node $R_m^{(\ell)}$ (i.e., a terminal node after the ℓ-th split), the best split is chosen according to the combination, say (x_j^*, s_j^*, R_m^*), maximizing the effect size $f^{(\ell)}$. The highest effect size determines the highest relative decrease in deviance when moving from a more parsimonious model to a less parsimonious one. Trunk-growing proceeds until the size of each terminal node of the trunk is higher than a user-defined minimum size or, alternatively, until the maximum number of splits L is reached (L is also defined by the user).

Once the trunk-growing is complete, pruning is carried out using CART-like V-fold cross-validation. The 'best' size of the trunk corresponds to the one minimizing the cross-validated prediction accuracy as well as its standard error. Likewise in CART, a '$c \cdot SE$' rule is used, where c is a constant. Simulation results presented in Dusseldorp et al. (2010) for RTA suggest to set $c = 0.80$ if $n \leq 300$ and $c = 0.5$ otherwise.

Pruning is a fundamental step of the algorithm since the number of terminal nodes of the trunk and their relative split points, as well as the splitting predictors, determine the number, order and type of threshold interactions terms to be included in the trunk model. Taking advantage of the extreme flexibility of the recursive partitioning algorithm, STIMA is applicable to all types of predictors (both numeric

and nominal) and it can be extended to model all types of interactions (nominal by numeric, nominal by nominal, numeric by numeric).

3 Model Selection Issues

The CTA and RTA methods can be seen as a variable selection criteria: the most influential variables are chosen as splitting predictors in the trunk-growing process of STIMA. Consequently, the trunk-pruning process of STIMA can be seen as a model selection procedure since it allows to retain in the model the most important interaction effects only. Since the trunk is derived by recursively partitioning the original data, the main role in the definition of the interaction effect is played by the splitting predictors, particularly the first one, which concur to define the trunk.

3.1 Current Implementation

One of the user-defined options currently implemented in STIMA is the possibility to force the first split of the trunk by an a-priori selection of the first splitting predictor. This manual selection of the first split can be motivated by research questions from social sciences concerning moderator effects of some nominal predictor. It allows to specifically account for the effect of a treatment factor affecting in a different way subsets of observations (see Dusseldorp et al. 2007 for an example).

Alternatively, when the analysis is not aimed to the evaluation of the influence of such a moderator effect, the selection of the first split of the trunk is performed automatically. The current implementation of STIMA is based on the a-priori estimation of a sequence of trunks for each single predictor j ($j = 1, \ldots, J$). Each sequence is composed of L trunks presenting 1 up to L splits. For each of these trunks, the evaluation of the predictive accuracy is made by performing V-fold cross-validation. As a result, the first splitting predictor is the one whose sequence provides a trunk with the lowest cross validation error among those provided by the $J \times L$ trunks.

Once the first splitting predictor has been chosen, the algorithm re-estimates the trunk sequence and re-computes the V-fold cross validation errors as well as their standard errors. The selection of the final model is performed according to the '$c \cdot SE$' rule as specified in Sect. 2.2. In addition, backward variable selection can be performed in order to further reduce model complexity.

3.2 Alternative Approach

As a matter of fact, the previously described model selection procedure does not necessarily provide the right compromise between the accuracy and the parsimony

of the selected model. To investigate this point, an alternative approach is hereby presented. It is based on the estimation of a set of plausible trunk models for the analyzing data and on the subsequent selection of the most appropriate one. It consists of the following steps:

1. For each predictor j ($j = 1, \ldots, J$), estimate L trunk models ($L > 1$) by using STIMA with j as the first splitting predictor and ℓ interaction terms ($\ell = 1, \ldots, \ell = L$).
2. For each estimated model, use the backward selection option to reduce its size $k_{(j,\ell)}$ ($1 \leq k_{(j,\ell)} \leq J + L - 1$), which corresponds to the number of model parameters.
3. For each of the $J \times L$ estimated models, retain j and $k_{(j,\ell)}$ and compute the goodness of fit of the model $f_{j,k_{(j,\ell)}}$ and its generalization error $e_{j,k_{(j,\ell)}}$.
4. Repeat B times Steps 1–4 by bootstrapping the original data.
5. For each j and for each $k_{(j,\ell)}$, average $k_{(j,\ell)}$, $f_{j,k_{(j,\ell)}}$ and $e_{j,k_{(j,\ell)}}$ over the B runs, such to obtain $\bar{k}_{(j,\ell)}$, $\bar{f}_{j,\bar{k}_{(j,\ell)}}$ and $\bar{e}_{j,\bar{k}_{(j,\ell)}}$, respectively.
6. Set: $\bar{k}_{min} = \min(\bar{k}_{(j,\ell)})$ and $\bar{k}_{max} = \max(\bar{k}_{(j,\ell)})$. Define the subset $\mathscr{M}^*$ of $(\bar{k}_{max} - \bar{k}_{min} + 1)$ trunk models: each model $m_{\bar{k}}$ in $\mathscr{M}^*$ corresponds, for a given $\bar{k}_{(j,\ell)}$, to the model presenting the lowest value of the average generalization error $\bar{e}_{j,\bar{k}_{(j,\ell)}}$, namely: $e^*_{\bar{k}} = \arg\min_{\bar{k}}(\bar{e}_{j,\bar{k}_{(j,\ell)}})$.
7. For each consecutive pairs of models $m_{\bar{k}}$ and $m_{\bar{k}+1}$ ($\bar{k} = 1, \ldots, \bar{k}_{max} - 1$) in $\mathscr{M}^*$, compute an heuristic measure $T_{m_{\bar{k}},m_{\bar{k}+1}}$ (to be defined later) which evaluates the *trade-off* between the change in model accuracy and the associated change in model fitting when passing from a most parsimonious model ($m_{\bar{k}}$) to a less parsimonious one ($m_{\bar{k}+1}$).
8. Select $\hat{m}_{\bar{k}} \in \mathscr{M}^*$ as the less parsiomonious model such that $T_{m_{\bar{k}},m_{\bar{k}+1}} \geq 1$.

From a practical point of view, the proposed approach allows to perform model selection in STIMA by simultaneously considering, in each bootstrap replication, model fitting and model accuracy. The first one is evaluated by computing one of the information criteria typically used in the GLM framework such as, for example, Akaike's AIC. The second one is evaluated by considering the generalization error of each estimated model through, for example, test set data or V-fold cross-validation. In addition, the outcomes deriving from each bootstrap replication are averaged in order to robustify the results as well as to reduce the possible influence of outliers and to account for the problem of the variable selection bias which typically affects recursive partitioning algorithms.

As stated above, the selection of the final model $\hat{m}_{\bar{k}} \in \mathscr{M}^*$ derives from an heuristic measure $T_{m_{\bar{k}},m_{\bar{k}+1}}$, which evaluates the trade-off between model fitting and accuracy. To define it, the subset of models $\mathscr{M}^*$ as specified in Step 6. on the basis of the minimum values of the average generalization error for each model size ($e^*_{\bar{k}_{\min}}, \ldots, e^*_{\bar{k}}, \ldots, e^*_{\bar{k}_{\max}}$) is considered. Since information criteria used in GLM usually decrease as long as accuracy increases, it often happens that the average goodness of fit measure $\bar{f}_{j,\bar{k}_{(j,\ell)}}$ for each model in $\mathscr{M}^*$ also decreases when passing from $e^*_{\bar{k}_{\min}}$ to $e^*_{\bar{k}_{\max}}$.

The computation of $T_{m_{\bar{k}},m_{\bar{k}+1}}$ assumes that either $e^*_{\bar{k}}$ or its associated $\bar{f}_{j,\bar{k}_{(j,\ell)}}$ ranges in (0, 1). This involves that each $e^*_{\bar{k}}\left(\bar{f}_{j,\bar{k}_{(j,\ell)}}\right)$ is normalized to $\tilde{e}^*_{\bar{k}}\left(\tilde{f}_{j,\bar{k}_{(j,\ell)}}\right)$ as follows:

$$\tilde{e}^*_{\bar{k}} = \frac{e^*_{\bar{k}} - \min(e^*_{\bar{k}})}{\max(e^*_{\bar{k}}) - \min(e^*_{\bar{k}})}; \quad \tilde{f}_{j,\bar{k}_{(j,\ell)}} = \frac{\bar{f}_{j,\bar{k}_{(j,\ell)}} - \min(\bar{f}_{j,\bar{k}_{(j,\ell)}})}{\max(\bar{f}_{j,\bar{k}_{(j,\ell)}}) - \min(\bar{f}_{j,\bar{k}_{(j,\ell)}})}$$

For two consecutive models $m_{\bar{k}}$ and $m_{\bar{k}+1}$ in $\mathscr{M}^*$, $T_{m_{\bar{k}},m_{\bar{k}+1}}$ corresponds to:

$$T_{m_{\bar{k}},m_{\bar{k}+1}} = \frac{\tilde{e}^*_{\bar{k}} - \tilde{e}^*_{\bar{k}+1}}{\tilde{f}_{j,\bar{k}_{(j,\ell)}} - \tilde{f}_{j,\bar{k}+1_{(j,\ell)}}} \quad \forall \bar{k}, \bar{k}+1 \in (\bar{k}_{\min}, \ldots, \bar{k}_{\max} - 1)$$

$T_{m_{\bar{k}},m_{\bar{k}+1}}$ is computed sequentially, from the first (less accurate) model in $\mathscr{M}^*$ to the last but one (most accurate). The proposed criteria selects the final model $\hat{m}_{\bar{k}}$ as the most accurate model such that $T_{m_{\bar{k}},m_{\bar{k}+1}} \geq 1$. This corresponds to proceed sequentially in model selection as long as the increase in average model accuracy is greater than (or equal to) the increase in average goodness of fit.

4 Empirical Evidence

In the following, two examples on real data available on the UCI ML Repository (Asuncion and Newman 2007) are presented in order to evaluate the performance of the proposed model selection criteria when applied on STIMA. The first example involves the Liver Disorders dataset: the goal is to classify cases on the basis of two levels of alcohol consumption corresponding to the proportion of people presenting high alcohol consumption with respect to those classified as regular drinkers. Six numerical predictors are observed on 345 individuals. The second example involves a regression problem: the Housing dataset is analyzed in order to estimate the median value of 506 owner-occupied homes localized in different census tracts of Boston on the basis of 14 numeric and 1 binary predictors.

Figure 1 illustrates the model selection process as defined in Sect. 3.2 for the first dataset. The subset of models $\mathscr{M}^*$ is identified in the left panel: it is composed by those models that, for a given size $\bar{k}_{(j,\ell)}$, presents the lowest value of $\bar{e}_{j,\bar{k}_{(j,\ell)}}$ (i.e., models with 6, 7, 10 and 12 parameters, respectively). The right panel shows, for those models, the decrease in both $\tilde{e}^*_{\bar{k}}$ and $\tilde{f}_{j,\bar{k}_{(j,\ell)}}$ when passing from a more parsimoniuos model to a less parsimonious one. The selected model $\hat{m}_{\bar{k}}$ is the one with size $\bar{k}_{(j,\ell)} = 10$, since the decrease in $\tilde{e}^*_{\bar{k}}$ (0.37) is greater than that of $f_{j,\bar{k}_{(j,\ell)}}$ (0.33) when passing from $\bar{k}_{(j,\ell)} = 6$ to $\bar{k}_{(j,\ell)} = 10$. As a result, $T_{6,10} = 1.12 > 1$.

For both datsets, the performance of STIMA deriving from the standard implementation of model selection as well as from that proposed in this paper is compared

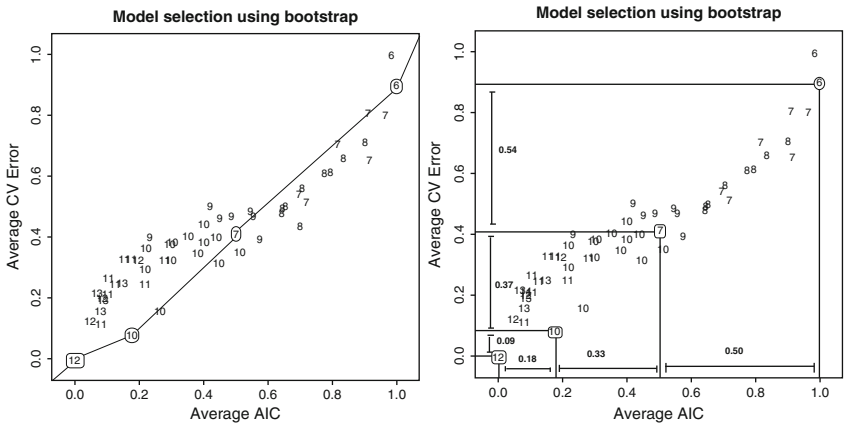


Fig. 1 Results of the model selection process described in Sect. 3.2 for the Liver Disorders dataset. The two plots show the relationship between $\widetilde{f}_{j,\bar{k}_{(j,\ell)}}$ and $\widetilde{e}^{*}_{\bar{k}}$ for the $J \times L$ classification trunk models estimated with $B = 50$ bootstrap replications. Each number in the plot refers to the average model size $\bar{k}_{(j,\ell)}$, i.e., the average number of parameters obtained over the B runs

Table 1 Benchmarking STIMA[a]

Model	*Stepwise GLM*	*Stepwise GAM*	*CART*	*MARS*	*STIMA*	$STIMA_2$
		Liver Disorders dataset				
CV Error	.287	.255	.287	.307	.249	.194
Standard Error	.024	.023	.024	.025	.024	.023
# parameters	11	21	2	9	8	10
		Housing dataset				
CV Error	.324	.197	.243	.167	.150	.116
Standard Error	.058	.035	.036	.028	.019	.014
# parameters	22	20	6	21	23	12

[a]The Table reports the 10-fold cross-validation error, its standard error and the number of parameters of each model. Compared models are GLM and GAM with stepwise variable selection and the manual search of interaction terms, CART, MARS and STIMA. As for the latter, STIMA refers to the standard implementation of the model selection process; $STIMA_2$ refers to the criteria presented in Sect. 3.2.

with GLM and Generalized Additive Models (GAM) (Hastie and Tibshirani 1990), both estimated with stepwise variable selection and with the manual search of cross-product interactions, as well as with that of CART and Multivariate Adaptive Regression Splines (MARS) (Friedman 1991). For each method cross-validation is used to estimate prediction accuracy. Results of the comparative analysis are summarized in Table 1: they show that the model identified by STIMA with the proposed model selection criteria ($STIMA_2$ in Table 1) overperforms its competitors in terms of cross-validation error without requiring a consistent additional number of parameters.

5 Concluding Remarks

The possible use of STIMA in statistical modeling is motivated by practical considerations about the search of interactions in GLM. When dealing with many predictors, as well as when no a-priori hypothesis about possible relationships among them can be formulated, the search of interactions is not straightforward but it is a quite complicated task. One impractical and time-consuming approach is testing all possible interactions and retain the most important ones. More straightforwardly, the trunk model is a suitable choice as it allows to automatically estimate the number, the order and the types of interactions. The proposed approach to model selection in STIMA is able to strenghten the effectiveness of the method, either in the classification or regression case, since it improves the accuracy of the estimated trunk model while keeping it more parsimonious.

Acknowledgements The author thanks the anonymous referee and Elise Dusseldorp for their helpful and valuable suggestions, which improved the overall quality of the paper. This research is supported by the research funds awarded by University of Cagliari within the 'Young Researchers Start-Up Programme 2007'.

References

Asuncion, A., & Newman, D. J. (2007). UCI machine learning repository, http://archive.ics.uci.edu/ml/.

Berrington de Gonzalez, A., & Cox, D. R. (2007). Interpretation of interaction: A review. *Annals of Applied Statistics*, *1*(2), 371–375.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Conversano, C., & Dusseldorp, E. (2010). Simultaneous threshold interaction detection in binary classification. In C. N. Lauro, M. J. Greenacre, & F. Palumbo (Eds.), *Studies in classification, data analysis, and knowledge organization* (pp. 225–232). Berlin-Heidelberg: Springer.

Dusseldorp, E., Conversano, C., Van Os, B. J. (2010). Combining an additive and tree-based regression model simultaneously: STIMA. *Journal of Computational and Graphical Statistics, 19*(3), 514–530.

Dusseldorp, E., & Meulman, J. (2004). The regression trunk approach to discover treatment covariate interactions. *Psychometrika*, *69*, 355–374.

Dusseldorp, E., Spinhoven, P., Bakker, A., Van Dyck, R., & Van Balkom, A. J. L. M. (2007). Which panic disorder patients benefit from which treatment: Cognitive therapy or antidepressants? *Psychotherapy and Psychosomatics*, *76*, 154–161.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, *19*, 1–141.

Hastie, T. J., Tibshirani, R. J., & Friedman, J. H. (2001). *Elements of statistical learning*. NewYork: Springer.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. London, New York: Chapman and Hall.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, *135*, 370–384.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London: Chapman & Hall.

Conditional Classification Trees by Weighting the Gini Impurity Measure

Antonio D'Ambrosio and Valerio A. Tutore

Abstract This paper introduces the concept of the conditional impurity in the framework of tree-based models in order to deal with the analysis of three-way data, where a response variable and a set of predictors are measured on a sample of objects in different occasions. The conditional impurity in the definition of splitting criterion is defined as a classical impurity measure weighted by a predictability index.

1 Introduction

Classification and regression trees have become a fundamental approach to data mining and prediction (Hastie et al. 2001). Non-parametrical methods based on classification and regression tree procedures can be fruitfully employed to handle the problem of the analysis of large datasets characterized by high-dimensionality and nonstandard structure.
The information in the data can be summarized in two (not exclusive) purposes:

- to investigate the structure of the data through the analysis and the interpretation of *exploratory trees*;
- to produce accurate classifiers/predictors in the form of *decision trees* to be used on new cases.

In this paper we focalize our attention on classification tree as explorative tool for three-way data.

2 Our Proposal

As a matter of fact, dealing with complex relations among the variables, every CART-based approach offers unstable and not interpretable solutions. This paper aims to define a segmentation methodology for three-way data matrix starting from

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_31,

some recent results. In fact, some contributions in this framework, following two-stage philosophy (Mola and Siciliano 1992), have been proposed by Tutore et al. in the recent past (Tutore et al. 2007).

2.1 The Data

The three ways of the dataset are cases, attributes and situations, respectively. Let **D** be the three-way data matrix of dimensions N, V, K, where N is the number of cases, objects or units, V is the number of variables, K is the number of situations. Assume that the V variables can be distinguished into two groups, namely there are M predictor variables $X_1, \ldots, X_m, \ldots, X_M$ and C response variables $Y_1, \ldots, Y_c, \ldots, Y_C$ where $M + C = V$. The K situations refer to modalities of a stratifying variable, which is called *instrumental variable*. Alternatively a time variable can be also considered for longitudinal data analysis. Predictors can be of categorical and/or numerical type whereas responses can be either categorical or numerical, thus a distinction can be made between a classification problem and a regression problem respectively. In the following we consider the case in which $C = 1$ and the response variable is categorical.

2.2 The Previous Work

Tutore et al. (2007) introduced a splitting criterion that considers the use of an instrumental variable called Partial Predictability Trees which is based on the two-stage splitting criterion (Mola and Siciliano 1992) and on the predictability τ index of Goodman and Kruskal (1954, 1979) for two-way cross-classifications. Two-stage algorithm works as follow: in the first stage, the best predictor is found maximizing the global prediction with respect to the response variable; in the second stage, the best split of the best predictor is found maximizing the local prediction. It can be proved that skipping the first stage maximizing the simple τ index is equivalent to maximizing the decrease of impurity in CART approach. This criterion was extended in Tutore et al. (2007) in order to consider the predictability power explained by each predictor with respect to the response variable conditioned by the instrumental variable Z. For that, the multiple τ_m and the partial τ_p predictability indexes of Gray and Williams (1981) have been considered.

At each node, in the first stage, among all available predictors X_m for $m = 1, \ldots, M$, the partial index $\tau_p(Y, X_m|Z)$ is maximized to find the best predictor X^* conditioned by the instrumental variable Z:

$$\tau_p(Y, X_m|Z) = \frac{\tau_m(Y|X_m, Z) - \tau_s(Y|Z)}{1 - \tau_s(Y|Z)} \tag{1}$$

where $\tau_m(Y|X_m, Z)$ and $\tau_s(Y|Z)$ are the multiple and the simple predictability measures. In the second stage, the best split s^* of the best predictor X^* is found by maximizing the partial index $\tau_s(Y|s, Z)$.

2.3 The Conditional Impurity

A partitioning algorithm can be understood as a recursive and binary segmentation of N objects into two groups such to obtain internally homogeneous and externally heterogeneous subgroups with respect to a response variable. At any internal node of the tree, a predictor generates the splitting variable (i.e., dummy variable) to discriminate the objects falling into the left subnode from those falling into the right subnode. Let s be the splitting variable generated by the mth predictor; let Q be the set of all the splitting variables generated by the mth predictor; let $i, \ldots, I$, $j = 1, \ldots, J$ and $k = 1, \ldots, K$ are respectively the subscripts for the ith category of the response variable, the jth category of the mth predictor and kth category of the stratifying variable; let $\frac{np}{N}$ be the proportion of cases in a generic parent node, $\frac{nl}{N}$ be the proportion of cases in the left child node and $\frac{nr}{N}$ be the proportion of cases in the right child node.

According to the CART splitting criterion (Breiman et al. 1984), the following decrease of impurity is maximized for each s ∈ Q

$$\Delta i\,(s,t) = i\,(t)\,\frac{np}{N} - i\,(tl)\,\frac{nl}{N} - i\,(tr)\frac{nr}{N} \tag{2}$$

where $i\,(\cdot)$ is the impurity function. It is clear that if a generic node is maximally pure (i.e., the impurity is equal to zero), none splitting rule can be defined because that node is obviously a perfect terminal one. Depending on the definition of the impurity function, different rules can be used such as example for the Gini index of heterogeneity.

We consider the predictability τ index of Goodman and Kruskal (1954) for two-way cross-classifications:

$$\tau_{Y|Z}\,(t) = \frac{\sum_{ik} f_{ik}^2 / f_{.k} - \sum_i f_{i.}^2}{1 - \sum_i f_{i.}^2} \tag{3}$$

where f_{ik} is the proportion of cases that belongs to class i of the response variable and that have category k of Z at a generic node, and $f_{.k}$ is the proportion of cases that have category k of Z at the same node.

The τ of Goodman and Kruskal is an index that considers the predictive strength of a generic Z variable on a Y variable in a two-way table. We propose a combination of a classical impurity measure such as the Gini diversity index with the last index in this way:

$$i\,(t) = \left(1 - \sum_i f_i^2\right)\left(1 - \left[\frac{\sum_{ik} f_{ik}^2/f_{.k} - \sum_i f_{i.}^2}{1 - \sum_i f_{i.}^2}\right]\right) = 1 - \sum_{ik} f_{ik}^2/f_{.k} \quad (4)$$

As it is well known, the τ of Goodman and Kruskal is defined between 0 and 1. The more it is close to 1, the more the Z variable explains the distribution of Y given Z; then, it is easy to verify that the more the τ is close to 0, the more the impurity measure is similar to the one used with classical CART.

We call this kind of impurity measure *conditional impurity* because it is a weighted average of each partial impurity in every class of Z stratifying variable. In this case the total impurity of Y variable is corrected by a factor which is complement to one of predictive strength of Z on Y. If Z and Y are independent we have exactly the same results of CART. So conditional impurity measure tries to find the best compromise between the unconditional impurity (simple Gini's diversity index) and the internal impurity explained by the measure $1 - \tau$.

As $\frac{nl}{N} + \frac{nr}{N} = \frac{np}{N}$, we can substitute $\frac{nl}{N}$ with p_{tl} and $\frac{nr}{N}$ with p_{tr}, so we can say that, in relation to the sample size of any generic parent node, $p_{tl} + p_{tr} = 1$ because starting by any parent node always we have that $p_{tl} + p_{tr} = \frac{np}{N}$. Following this approach, we define the decrease of impurity at generic node t as

$$\begin{aligned}\Delta i\,(s\,|t, Z) &= 1 - \sum_{ik} f_{ik}^2/f_{.k} - \left(1 - \sum_{ik|s=0} f_{ik}^2/f_{.k}\right) p_{tl} - \left(1 - \sum_{ik|s=1} f_{ik}^2/f_{.k}\right) p_{tr} \\ &= \sum_{ik|s=0} f_{ik}^2/f_{.k}\,(p_{tl}) + \sum_{ik|s=1} f_{ik}^2/f_{.k}\,(p_{tr}) - \sum_{ik} f_{ik}^2/f_{.k} \quad (5)\end{aligned}$$

in which the categories of the splitting variable s are denoted respectively by $s = 0$ and $s = 1$, by recalling that every splitting variable generated by the mth predictor is always binary.

3 Application Studies

We present two different sets of data to show how our proposal works: one from a simulation study and another one from a real dataset.

3.1 Simulated Dataset

Simulation study has been defined thinking to reliable situations in which our proposal can be functional. For that reason, a simulated dataset (Table 1) was built using different random distributions for the set of common variables (Discrete uniform, Normal, Binomial), whereas the response variable was generated by a non linear link with other variables (namely, X_1, X_3 and X_5). Variables X_2 and X_4 play the

Table 1 Simulations setting

Predictors		Binary response variable
X_1	Uniform in 1, 4	$y\ sin(k + X_5\pi) + 0.8X_3 - 0.1X_1X_3 + \epsilon$
X_2	Uniform in 1, 10	Y 1 if $y > 0$, 0 otherwise
X_3	Normal standard	
X_4	Binomial(10, 0.6)	
X_5	Binomial(3, 0.7)	
X_6	Binomial(1, 0.5)	

role of masking variables. To stress the methodology ability to explain hierarchical structure of data, different conditional variable Z, with different strength, have been simulated.

For each value of the stratifying variable both the CART and the Conditional Classification Tree were built. CART was considered both including and not including the instrumental variable in the set of the predictors. Four indexes are computed for both the classical CART and the Conditional Classification Tree:

- M.R.T., namely the *Misclassification ratio tree*. It is simply the error of the tree without taking in account the role of the Z (weighted by the number of cases in each of its categories);
- S.M.R. root, namely the *Stratified misclassification ratio at the root node*. It shows the error rate at the root node taking in account the role of the Z;
- S.M.R.T., namely the *Stratified misclassification ratio of the tree*. It shows the error rate of the tree model weighted by the Z;
- *Gain in accuracy*. It shows how the final tree improves in respect to the trivial situation (root node).

In Table 2, column named CART^a indicates the classifier that does not include the stratifying variable as predictor, as well as the column called CART^b indicates the classifier that includes the Z variable as predictor. In the table there are three main rows: the first one is relative to a strong relationship between Z and Y ($\tau_{Y|Z} = 0.6630$), the second one is characterized by a weak relationship ($\tau_{Y|Z} = 0.0142$) whereas in the third row Z and Y are independent. The results shown in the table are validated through V-fold cross-validation procedure. In the first case the conditional final stratified misclassification ratio is smaller than CART one; in this case the instrumental variable plays a strong role in determining the conditional split. In the second case the conditional stratified misclassification ratio is weakly lower than the other one: therefore, if there is not an a priori strong relationship, conditional tree works better than classical CART. In the third case the results are exactly the same: in fact Z and Y are independent.

3.2 Italian Bank Credit Dataset

We show some results using a survey led by an Italian credit bank. Table 3 presents the structure of the dataset.

Table 2 Simulated data: Main results

Simulated dataset	Conditional miscl. ratio	CART^a miscl. ratio	CART^b miscl. ratio	Stratifying strength
M.R.T.	0.0740	0.1291	0.0983	
S.M.R. root	0.3496	0.3496	0.3496	$\tau_{Y\|Z}$
S.M.R.T.	0.0527	0.1930	–	0.6630
Gain in accuracy	0.8493	0.4479	–	
M.R.T.	0.0860	0.1591	0,1935	
S.M.R. root	0.4218	0.4218	0.4218	$\tau_{Y\|Z}$
S.M.R.T.	0.2031	0.2272	–	0.0142
Gain in accuracy	0.5185	0.4614	–	
M.R.T.	0.0956	0.0956	0.0956	
S.M.R. root	0.4371	0.4371	0.4371	$\tau_{Y\|Z}$
S.M.R.T.	0.2411	0.2411	–	0.0000
Gain in accuracy	0.4484	0.4484	–	

Table 3 Italian bank credit dataset

Dataset variables	
1. Account Status (Stratifying Variable)	10. Present residence since
2. Other building owner	11. Education level
3. Purpose	12. Age
4. Duration	13. Credit history
5. Saving accounts/bond	14. Housing
6. Present employment since	15. N. of Bank Accounts
7. Personal status	16. Job
8. Gender	17. People being liable to provide maintenance for
9. Others debtors/guarantors	18. Existing credits at this bank
Response:	Type of Client (Good/Bad)

We choose as instrumental variable the account status; the response variable is the type of client (good or bad client). Table 4 shows the difference in badness of fit indexes of both classical CART (both including (CART^b) and not including (CART^a) the Z variable in the set of predictors) and the Conditional Tree. The strength of the relationship between Y and Z is equal to 0.1237. All the classifiers were validated through V-fold cross-validation.

The Fig. 1 shows an example of how each terminal node of the Conditional Classification Tree can be interpreted. For each level of stratifying variable a bar chart of the two categories of the Y variable is shown. In this terminal node there are 173 individuals about which $52,02\%$ are good clients and $47,98\%$ bad clients, so the error rate of this node is equal to 0.4798 and the assignment rule is *good client*. Nevertheless, looking at the figure, it can be noted that for two levels of the Z variable (in particular for the first and the second level of the Z), the mode is *bad client*. The error rate of this node must be weighted by the number of individuals belonging to each category of the stratifying variable. In this case, it is equal to 0.3006.

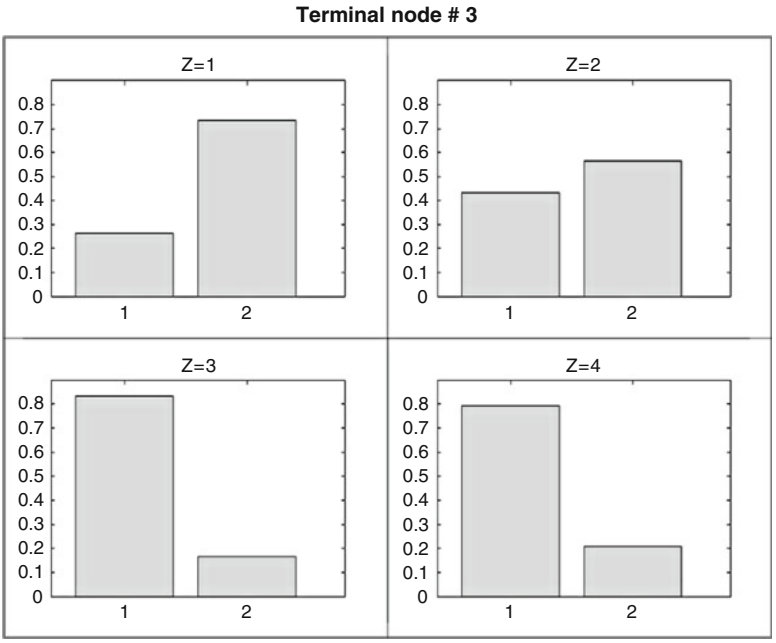


Fig. 1 Zoom on the terminal node # 3-Italian bank dataset

Table 4 Italian bank credit: Main results

Italian bank dataset	Conditional misscl. ratio	CART^a misscl. ratio	CART^b misscl. ratio	Stratifying strength
M.R.T	0.2770	0.2740	0.2310	
S.M.R. root	0.3000	0.3000	0.3000	$\tau_{Y\|Z}$
S.M.R.T	0.2180	0.2420	–	0.1237
Gain in accuracy	0.2733	0.1933	–	

Table 4 shows the badness of fit of the built model compared with the ones which are returned by the classical CART. Note that the Misclassification Ratio of the Conditional Tree *without* take in account the role of the stratifying variable is higher than the same index of the classical CART (with and without the instrumental variable included as predictor). The role played by the stratifying variable in governing the splitting rule can be appreciated looking at the Stratified Misclassification Ratio of the tree. In this case this index is lower than the one of the CART^a, as confirmed also by its higher gain in accuracy.

This example on a real data set shows how, even if we have not an important improvement in terms of accuracy (error rates), it is possible to get a better interpretation of the phenomena under study. Indeed, including the stratifying variable in the CART we obtain a M.R.T. index really close to the one of the Conditional Tree (see the column CART^b in the Table 4). In this case the first split (the split at the root node) was governed by the Z variable, and never that variable generated other splits

in the building of the tree. When it happens, the role played by the stratifying variable is not interpretable. In other words, our goal is not to improve the performance of the model in terms of decrease of error rates, but to improve the performance of the model in terms of its interpretability without losing in accuracy.

4 Concluding Remarks

In this paper we have proposed a different splitting criterion for classification trees. When data are characterized by a stratified structure, we choose to weight a classical impurity measure for a quantity that takes in account how the stratifying variable can explain the response variable. In this case the instrumental variable does not contribute to determine the split in a direct way, but its contribution is 'more than direct' in the sense that it is present into the determination of all splits. From the explorative point of view, such a tree-based model allows the interpretation of all the categories of the response variable taking into account all the classes of the instrumental variable with only one tree. What we present is neither better nor worse than a classical approach such as the CART methodology. In this work, our intention is the introduction to a new way to interpret the splitting criterion with this kind of data.

Acknowledgements Authors wish to thank anonymous referees for their helpful comments.

References

Breiman, L., Friedman, J. H., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Wadsworth, Belmont, California.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross-classification. *Journal of American Statistical Association*, *48*, 732–762.

Goodman, L. A., & Kruskal, W. H. (1979). *Measures of association for cross classifications*. New York: Springer-Verlag.

Gray, L.N., & Williams, J.S. (1981). Goodman and Kruskal's tau b: Multiple and partial analogs. *Sociological Methods & Research, 10*(1), 50–62.

Hastie, T. J., Tibshirani, R. J., & Friedman, J. (2001). *The elements of statistical learning*. New York: Springer-Verlag.

Mola, F., & Siciliano, R. (1992). A two-stage predictive splitting algorithm in binary segmentation. In Y. Dodge & J. Whittaker (Eds.), *Computational statistics: COMPSTAT '92* (Vol. 1, pp. 179–184). Heidelberg (D): Physica Verlag.

Tutore, V. A., Siciliano, R., & Aria, M. (2007). Conditional classification trees using instrumental variables. In *Proceedings of the 7th IDA2007 conference (Ljubljana, 6–8 September, 2007)*, Lecture Notes in Computer Science Series of Springer.

Part VII
Analysis of Financial Data

Visualizing and Exploring High Frequency Financial Data: Beanplot Time Series

Carlo Drago and Germana Scepi

Abstract In this paper we deal with the problem of visualizing and exploring specific time series such as high-frequency financial data. These data present unique features, absent in classical time series, which involve the necessity of searching and analysing an aggregate behaviour. Therefore, we define peculiar aggregated time series called beanplot time series. We show the advantages of using them instead of scalar time series when the data have a complex structure. Furthermore, we underline the interpretative proprieties of beanplot time series by comparing different types of aggregated time series. In particular, with simulated and real examples, we illustrate the different statistical performances of beanplot time series respect to boxplot time series.

1 Introduction

High-frequency financial data (Engle and Russell 2009) are observations on financial variables collected daily or at a finer time scale (such as time stamped transaction-by-transaction, tick-by-tick data, and so on). This type of data have been widely used to study various market microstructure related issues, including price discovery, competition among related markets, strategic behaviour of market participants, and modelling of real-time market dynamics. Moreover, high-frequency data are also useful for studying the statistical properties, volatility in particular, of asset returns at lower frequencies. The analysis of these data is complicated for different reasons. We deal with a huge number of observations (the average daily number of quotes in the USD/EUR spot market could easily exceed 20,000), often spaced irregularly over time, with diurnal patterns, price discreteness, and with a complex structure of dependence. The characteristics of these data don't allow for visualizing and exploring by means of the classical scalar time series. Furthermore, it is very difficult to forecast data without defining an aggregate behaviour.

In this paper we introduce beanplot time series (in Sect. 2) with the aim of synthesizing and visualizing high-frequency financial data or, more in general, complex types of temporal data. We discuss their properties by proposing critical

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_32,

comparisons among different possible aggregated time series (in Sect. 3). In particular, we have carried out several simulated examples (in Sect. 3.1), starting from different models, different number of observations and different intervals of aggregation to show how beanplot time series perform better than boxplot time series. Some interpretative rules are given in Sect. 3.2. We have enriched our analysis with an application on real high frequency financial data where we show how beanplot time series easily detect real intra-day patterns (in Sect. 3.3).

2 Beanplot Time Series

The Beanplot time series $\{b_{Y_t}\} t = 1 \ldots T$ is an ordered sequence of beanplots or densities over time. The time series values can be viewed as realizations of an X beanplot variable in the temporal space T, where t represents the single time interval. The choice of the length of the single time interval t (day, month, year) depends on the specific data features and objectives the analyst wants to study.

A beanplot realization at time t is a combination between a 1-d scatterplot and a density trace. It is defined (Kampstra 2008) as:

$$\hat{f}_{h,t} = \frac{1}{nh} \sum_{n}^{i=1} K(\frac{x - x_i}{h}) \tag{1}$$

where x_i $i = 1 \ldots n$ is the single observation in each t, K is a Kernel and a h is a smoothing parameter defined as a bandwidth.

It is possible to use various Kernel functions: uniform, triangle, epanechnikov, quartic (biweight), tricube (triweight), gaussian and cosine. The choice of the kernel, in the beanplot time series is not particularly relevant because our simulations show that the different kernels tend to fit similarly the underlying phenomena. Some differences reveal themselves in presence of outliers. In these cases a better kernel seems to be the Gaussian kernel which is more robust. Looking at the characteristics of our data, we have chosen this kernel for the different applications.

The choice of the h value is much more important than the choice of K (Silverman 1986). With small values of h, the estimate looks "wiggly" and spurious features are shown. On the contrary, high values of h give a too smooth or too biased estimate and it may not reveal structural features, as for example bimodality, of the underlying density. In literature, several methods for choosing the bandwith were proposed. Practically, each paper which proposes a new bandwidth selection method contains or reports a small simulation study which compares the performance of the new method with that of one or two other methods on two to four densities. However neither of these studies gave a clear answer which bandwidth selection method was the best. With a visualization aim, we use the Sheather-Jones criteria (1991) that defines the optimal h in a data-driven approach in our application.

The kernel density estimators can be compared with other non parametric methods of density estimation (Fryer 1977). Empirical results show that, for example, the

splines smooth out the original data. This implied that we lost some relevant data features. Therefore kernel density estimators seem very useful in explorative contexts while spline smoothers retain the very relevant data features, not taking into account some irregularities which however arise in the case of complex data such as high frequency data.

3 Different Aggregated Time Series for High Frequency Data

With the aim of summarizing and visualizing high frequency time series, different types of aggregated time series can be considered. The mean or the total of the single values represent weak aggregations because important information is neglected.

Initially, we used stripchart time series. This type of time series correctly shows the original trend as well as the minimum and the maximum of each interval (a day, for example). However in such graphics, one dot is plotted for each observation in the single time interval and, consequently, it is a useful tool only when there are very few points. Therefore, it might be difficult to apply them in the high frequency data framework.

A recent proposal (Arroyo 2009), in the context of symbolic data, consists in substituting time series of observations with histogram time series. Histograms are very useful for temporal and spatial aggregations for many reasons: their simple and flexible structure, their capacity to describe the essential features of the data with reasonable accuracy and their closeness to the data, without imposing any distribution. Nevertheless, the multiple histograms are difficult to compare when there are many of them plotted on a graph, because the space becomes cluttered.

Tukey's boxplot (1977) is commonly used for comparing distributions between groups. For time series data, the boxplot seems to show several features of the temporal aggregation: center, spread, asymmetry and outliers. Furthermore, box plot time series well detect the main structural changes. However, the number of outliers detected will increase if the number of observations grows and the information about the density is neglected. This information can be very important in the aggregation of high frequency financial data where different volatility clusters can arise. In order to retain this information, it is possible to use violin plot time series. This tool (Benjamini 1998) combines the advantages of boxplots with the visualization of the density and it provides a better indication of the shape of the distribution. However, in a violin plot the underlying distribution is visible but the individual points, besides the minimum and maximum, are not visible and not indication of the number of observations in each group is given.

Our proposal consists in using beanplot time series in the context of high frequency financial data. Indeed, in each single beanplot all the individual observations are visible as small lines in a one-dimensional scatter plot, as in a stripchart. In the beanplot time series, both the average for each time interval (represented by the beanline) and the overall average is drawn; this allows an easy comparison among temporal aggregations.

The estimated density of the distribution is visible and this demonstrates the existence of clusters in the data and highlights the peaks, valleys and bumps. Furthermore, anomalies in the data, such as bimodal distributions are easily spotted. This is very interesting information in the context of high frequency financial time series where the intra-period variability represents the main characteristics of the data. The number of bumps can be considered as a signal of different market phases in the daily market structure. We can also observe that the beanplot becomes longer in the presence of price anomalies such as peculiar market behaviours (speculative bubbles).

3.1 Experimental Evidence: A Simulation Study

In order to study the performance of beanplot time series in visualizing and exploring high frequency financial data, we conduct several experiments on different models. The experiments are designed to replicate different volatility processes (with increasing complexity). In this respect we study the capability of different aggregated time series (boxplot time series and beanplot time series) to capture the main features of the original data.

For our simulations we developed several algorithms in R (Ramsay and Silverman 2007). We generated 18 types of models, where each model represents a different univariate GARCH/APARCH time series model. In order to analyse the effect of the different number of observations on our results, we varied, for each model, the number of observations from 200,000 to 700,000. In this way, we simulated different types of financial markets. Initially, we decided to aggregate our data in ten different groups as ten different days. Then we have tested different time aggregations (by reducing or increasing the number of groups). Therefore in each day we had from 20,000 to 70,000 observations. Finally, to test our results we made 100 replications for each model.

The outcome for each computational experiment performed is the visualization of the different aggregated time series over the time. For each experiment we registered the captured statistical features from the beanplot time series compared to the original scalar time series and the boxplot time series.

The results of our simulations show in the first place that the beanplots tend to visualize a higher amount of information on the daily data, and in particular the intra-day patterns in the behaviour of the series, where the boxplots tend to return a smoothed view of the financial time series. We report here, by way of example, the results (in Fig. 1) with an underlying model (model 1) of the type GARCH(1,1) and those (in Fig. 2) obtained with a model (model 2) of the type AR(1,5-GARCH(1,1) both with 200,000 observations. By increasing the complexity of the time series we observe more clearly the differences between boxplot and beanplot time series. With beanplots, we are able to understand the structural changes and the different forms of the objects more distinctly. When the complexity reaches an extremely high level there is an increase of the outliers. Boxplot time series seem to suffer this higher

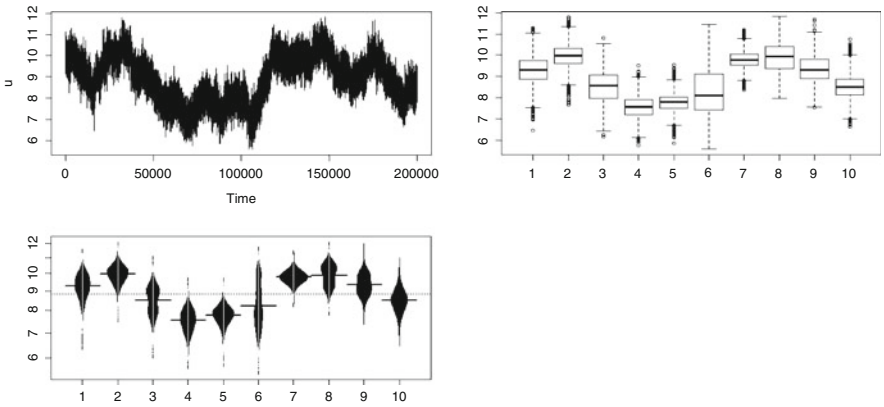


Fig. 1 Scalar, boxplot and beanplot time series for the model1

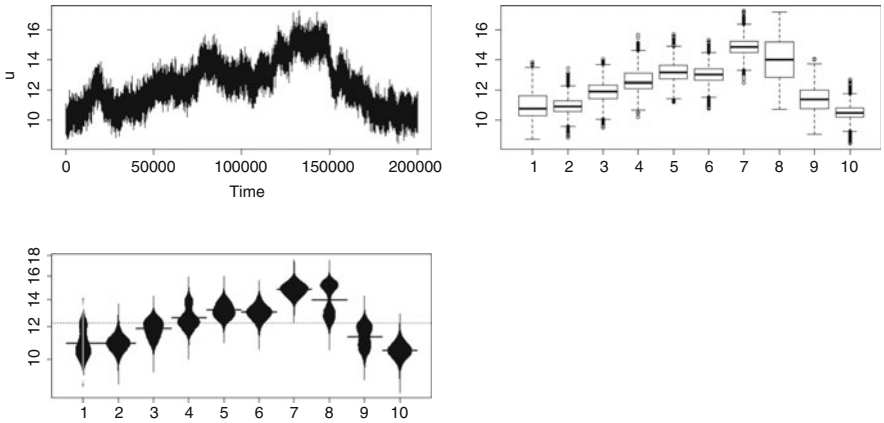


Fig. 2 Scalar, boxplot and beanplot time series for the model2

volatility of the markets. It is also interesting to note that by increasing the number of observations, beanplots alone may gave us clearer understanding and are then more useful than the boxplot time series. In fact, the number of outliers tends to increase and the boxplots become similar to each other (in Fig. 3 we report an example of the model1 with 700,000 observations). At the same time our simulations show that there is a specific number of observations that could be retained by choosing one interval or another. So, the choice of the interval seems to be linked to the interests of the applied researcher. In Fig. 4 we show the differences between beanplot time series with different temporal aggregations: a higher number of observations considered in the interval shows a higher number of bumps (and so of structural changes). The risk could be the loss of the information related to the cycles, where a lower number shows the structure of the series, but it is expensive in terms of space used (and there is the risk of not visualizing patterns).

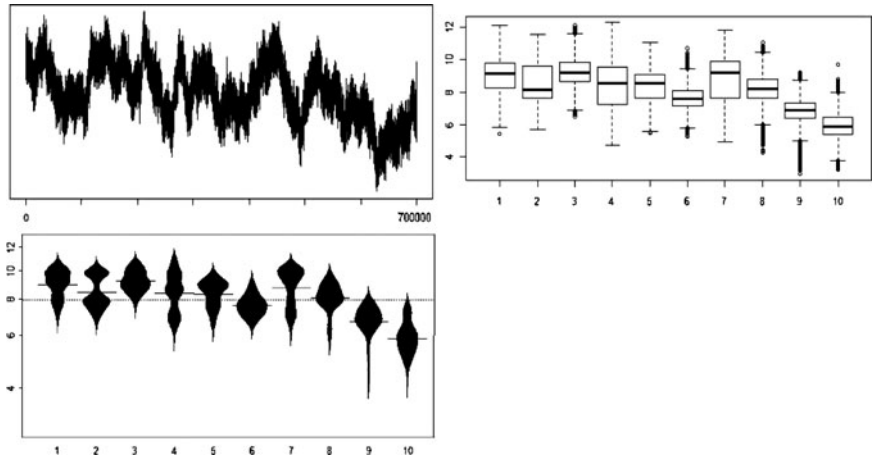


Fig. 3 Scalar, boxplot and beanplot time series for the model1 with 700,000 data

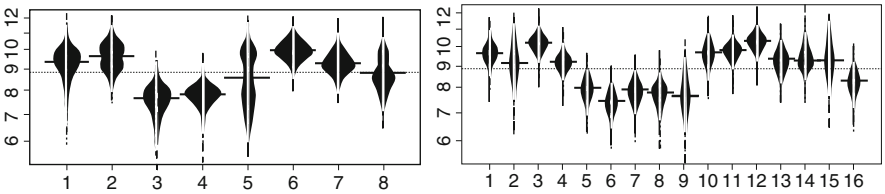


Fig. 4 Beanplot time series: different temporal aggregations

3.2 Some Empirical Rules

The aim of the analyst using the simulation data, is also to obtain some empirical rules. The most important characteristics of the beanplot time series is the capability to capture three relevant aspects in the dynamics of the complex data: the location, the size, the shape. Therefore, beanplot time series should be interpreted simultaneously considering this information.

The location shows the average or median price, and thus represents a useful benchmark for comparing different units. This parameter gives the possibility to visualize a time series trend. This feature is not possible with other smoothers or other nonparametric techniques, while in the beanplot time series we can explicitly consider a center for each time aggregation.

The size represents the general level of volatility, while the shape specifically represents the internal structure and the intra-day patterns. Therefore, by observing these parameters, we can easy identify speculative bubbles, structural changes, market crashes and so on. Furthermore, beanplot bumps can be seen as equilibrium values for the operators and they can be very important in trading strategies.

3.3 A Real Example on the Zivot Dataset

The data used in this application are contained in the Zivot dataset (Yan and Zivot 2003). These data are specifically related to the official TAQ (Trades and Quotes) database containing "tick by tick" data for all the stocks in NYSE from 1993. The Zivot dataset refers to 1997 and contains quotes and the trades for Microsoft. Here we consider the transaction prices for the period 1 May–15 May for a total of 11 days (except periods where the market is closed). Finally, we take into account 98,705 observations (instead of 98,724). In this case we do not consider the prices > 150, which allows us to avoid the data visualization completely. This exclusion does not modify the data structure.

In the original time series, we cannot read easily the intra-period variability because of the overwhelming number of observations. Therefore we represent boxplot and beanplot time series (Fig. 5) where the data are aggregated day by day. We observe over the different graphs the general view of the trend of the series, but also the variation (or the price volatility in a same day) of the same data. The outlier identification is straightforward and is imputable to a specific daily price change. However, we note that beanplot enriches the boxplot information by showing the intra-day structure. We can detect various structural changes in the period 1–8 May and the boxplot time series identify only major price changes. Each beanplot can be seen as the ideal "image" of the market at a specific time. In particular we can observe that the objects seem to be characterized by a response to the shocks, as the level (or the average) of the boxplots and the beanplots tends to change day by day. This phenomenon is due to the response of the time series to news that impose a different size, shape and location conditionally to the relevance of the shock. Changes in boxplot and beanplot levels seem to be directly influenced by daily news, where the number of bumps in the beanplot time series is directly linked to intra-day news. At the same time it is interesting also to note the volatility levels that seems to be

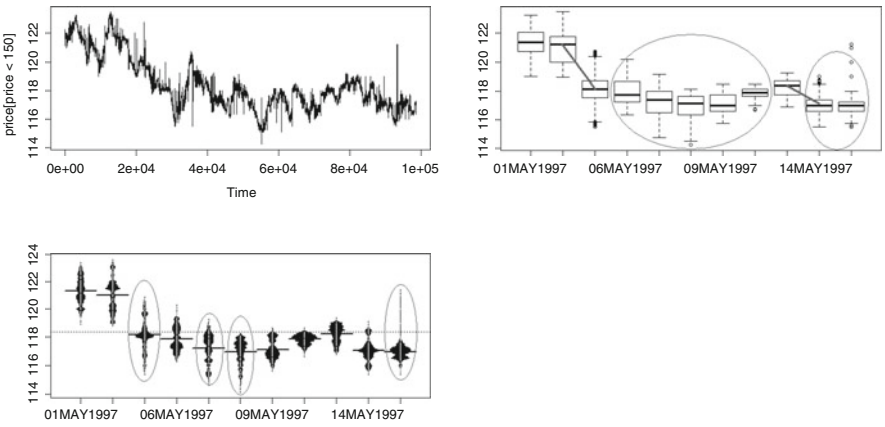


Fig. 5 Microsoft transaction prices series: boxplot and beanplot time series

higher after a single shock and tends to decrease over the time, and disappears after a few days. Finally, it is important to note that the structure of the time series appears highly irregular in the beanplot case. At the same time the boxplots tend to smooth the information contained in data, where the beanplots tend to reflect the complex behavior of the markets and the intra-daily patterns.

Various advances of beanplots time series can be considered, in particular in a multivariate framework. We will apply the beanplots in three different specific financial contexts: (a) monitoring of more than one stock at a time (b) pair trading using the cointegrated beanplot (c) risk analysis using the double beanplots. In particular, we are working further on the forecasting and the clustering of beanplot time series.

References

Arroyo, J., & Maté C. (2009). Forecasting histogram time series with k-nearest neighbours methods. *International Journal of Forecasting*, *25*, 192–207.

Benjamini, Y. (1988). Opening the box of the box plot. *The American Statistician*, *42*, 257–262.

Engle, R. F., & Russell, J. (2009). Analysis of High Frequency Data. In *Handbook of financial econometrics*, Vol.1 Tools and Techniques, North-Holland.

Fryer, M. J. (1977). A review of some non-parametric methods of density estimation. *Journal of the Institute of Mathematics Applications*, *20*, 335–354.

Kampstra, P. (2008). Beanplot: A boxplot alternative for visual comparison of distributions. *Journal of Statistical Software*, *28*(1), 1–9.

R Development Core Team. (2009). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B*, *53*, 683–690.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading: Addison-Wesley.

Yan, B., & Zivot G. (2003). *Analysis of high-frequency financial data with S-PLUS*. Working Paper. http://faculty.washington.edu/ezivot/ezresearch.htm.

Using Partial Least Squares Regression in Lifetime Analysis

Intissar Mdimagh and Salwa Benammou

Abstract The problem of collinearity among right-censored data is considered in multivariate linear regression by combining mean imputation and the Partial Least Squares (PLS) methods. The purpose of this paper is to investigate the performance of PLS regression when explanatory variables are strongly correlated financial ratios. It is shown that ignoring the presence of censoring in the data can cause a bias. The proposed methodology is applied to a data set describing the financial status of some small and medium-sized Tunisian firms. The derived model is interesting to be able to predict the lifetime of a firm until the occurrence of the failure event.

1 Introduction

A problem in multivariate linear regression commonly arises when two or more explanatory variables are strongly correlated. This phenomenon is well-known in statistics as multicollinearity. The regression parameter estimates may be unstable or even not computable. In lifetime analysis, this difficulty severely increases in the presence of right-censored data.

The problem raised by multicollinearity in regression can be treated by orthogonalization methods such as principal component regression (Massy 1965) and PLS regression (Tenenhaus 1998) or otherwise, by regularization-type methods such as ridge regression (Hoerl and Kennard 1970) or LASSO regression (Tibshirani 1996). The Partial Least Square methods are based on the NIPALS algorithm introduced by Wold (1966) in principal components analysis. In the last two decades, PLS methods have been used as a valuable tool in industrial applications, including marketing, sensorial analysis and genetics among others.

Censoring is another problem since information is incomplete and the time until event occurrence is unobserved. Recently, PLS regression was adapted to incorporate censored data in the Cox model leading to the PLS-Cox algorithm applied on microarray data (Bastien 2008). Datta et al. (2007) also proposed three ways to

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_33,

deal with right-censored data in PLS regression, namely the reweighting, the mean imputation and the multiple imputation approaches.

In this paper we discuss the use of the PLS regression in the assessment of the credit risk within a lifetime analysis framework. The application illustrated in the next sections, deals with a sample of small and medium-sized Tunisian firms, for which a response variable (the failure time) and a set of covariates based on accounting indicators, are collected over a number of years. The problem of right censoring is preliminarily dealt with the replacement of the unobserved response variable in the truncated observations, via a Mean imputation algorithm. Introducing a relatively new approach in the statistical analysis of financial distress, this paper aims to settle into a rather wide body of studies appeared over the last 40 years concerning the use of accounting indicators in the assessment of the failure risk: among the others, we cite for example the work of Altman (1968) who proposed the use of the discriminant analysis, Ohlson (1980) based on the employment of the logistic regression and Hillegeist et al. (2004) who have suggested a survival analysis approach.

The paper is organized as follows: A brief review of the lifetime analysis is provided in Sect. 2. The problem of treatment of censored data is outlined in Sect. 3. In Sect. 4, the detection of the multicollinearity among variables is illustrated using a data set composed of lifetime until failure occurrence and the financial ratios of Small and Medium-sized firms. Finally in Sect. 5, the PLS regression with right-censoring is explained and the obtained empirical results are discussed.

2 The Lifetime Analysis

The objective of lifetime analysis is to describe data that measure the time period until an event of interest occurs. For instance in financial applications, such an event may be the time of death or the occurrence of a critical event in the life of a given firm. In lifetime analysis, the right-censoring problem is encountered when the time of a failure event can not be observed. In fact, the time of failure is larger than some given fixed date delimiting the experimentation (Hougaard 2000). Here, we consider the lifetime as the duration from the birth time of the firm until the failure event for some small and medium-sized Tunisian firms. It is assumed that these firms are right-censored with respect to the year 2005.

In this paper, the lifetime is assumed to be a random positive real valued variable with a continuous distribution, independent with censoring. For censored data, a dummy variable δ is introduced, taking value 1 if the event is observed and 0 otherwise. In the lifetime analysis we are interested in estimating the survival function defined at any time point t as the probability of lifetime T being longer than t: $S(t) = P[T > t]$. The Kaplan and Meier (1958) estimator is a estimator of S defined as:

$$\hat{S}(t) = \prod_{T_i^{C'} < t} \left(1 - \frac{M(T_i^{C'})}{R(T_i^{C'})}\right) \tag{1}$$

where $T_i^C = min(T_i, C_i)$, with C_i and T_i are respectively the censoring and the uncensoring lifetime of observation i, $(T_1^{C'}, \ldots, T_i^{C'}, \ldots, T_n^{C'})$ are the $(T_1^C, \ldots, T_i^C, \ldots, T_n^C)$ classified by increasing order, $M(T_i^{C'})$ is the number of observed events to $T_i^{C'}$ and $R(T_i^{C'})$ is the number of individuals at risk before time $T_i^{C'}$.

3 Treatment of Censored Data

Given that the classical algorithms of the OLS regression and the PLS regression does not take into account the censoring issue, we use here the Mean imputation approach to deal with the right-censored data. The Mean imputation approach consists in keeping the observed response variable unchanged and replacing unobserved value by the quantity y_i^* expressed in terms of the Kaplan and Meier (1958) it is an estimator of the survival curve called Kaplan Meier estimator. Here we suppose that the maximal observed time t_{max} corresponds to an observed event ($\delta_{max} = 1$), even if $\delta_{max} = 0$. This hypothesis ensures that $\hat{S}(t_{max}) = 0$ (Datta et al. 2007). Formally, we define an imputed variable $\tilde{y}_i$ which satisfies:

$$\tilde{y}_i = \begin{cases} y_i \text{ if } \delta_i = 1 \\ y_i^* \text{ if } \delta_i = 0 \end{cases}$$

The determination of the quantity y_i^* is done via the following algorithm:
Let $t_1 < \ldots < t_{max}$ be the distinct event times,
Step 1: calculate $\Delta\hat{S}(t_j) = \hat{S}(t_{j-1}) - \hat{S}(t_j)$
Step 2: calculate $Z(t_j) = log(t_j)\Delta\hat{S}(t_j)$
Step 3: calculate $Z_i = \sum_{t_j > C_i} Z(t_j)$
Step 4: calculate $y_i^* = \frac{Z_i}{\hat{S}(C_i)}$

4 Detection of the Multicollinearity Among Variables

We consider a sample of 120 small and medium-sized Tunisian firms, a number of them have observed the failure event during the study period 1996–2007. These firms belong to different sectors of activity as shown in Table 1. Our data set is issued first from annual accounting reports of the firms: balance sheets, income statements, cash flow statements. The financial data related to right-censored firms are issued from the chartered accountants' offices and Tunisian national social security fund reports. The financial data relative to uncensored firms are sourced from registered auditors' reports, official balance sheet data and central bank of Tunisia reports.

We define the response variable Y as the lifetime until the failure event occurrence. The matrix X of explanatory variables is composed of 35 financial ratios describing the financial status of the firm 2 years before failure. Financial ratios

Table 1 Sectors of activity of the firms

Sector of activity	Right-censored firms	Uncensored firms	Total
Mechanics and Electronics (Industry)	7	7	14
Building Materials Ceramic and Glass (Industry)	9	7	16
Food-Processing (Industry)	8	4	12
Leather and Shoes (Industry)	9	8	17
Clothing (Industry)	8	5	13
Tourism (Service)	11	24	35
Building and Public Works (Service)	8	5	13

Table 2 Correlation between independent variables (Pearson correlation coefficients)[a]

(x_4, x_{10})	(x_7, x_8)	(x_7, x_{30})	(x_8, x_{30})	(x_{10}, x_{12})	(x_{15}, x_{16})	(x_{16}, x_{17})	(x_{17}, x_{34})	(x_{30}, x_{34})
0.658	0.682	0.612	0.670	0.695	0.597	0.649	0.608	0.655

[a]We have reported some correlation coefficients between variables having values higher than 0.5

given in Table 4 of the appendix consists in Financial Structure ratios $(x_1, \ldots, x_4)$, Profitability ratios $(x_5, \ldots, x_9)$, Liquidity ratios $(x_{10}, \ldots, x_{18})$, Solvency ratios $(x_{19}, \ldots, x_{23})$, Activity and Productivity ratios $(x_{24}, \ldots, x_{29})$, Growth ratios $(x_{30}, \ldots, x_{34})$ and Size firm ratio (x_{35}). It should be stressed here that financial ratios in statistical models may create poor model fits and/or predictions because the explanatory variables are correlated as shown in Table 2. Another way to asses the magnitude of multicollinearity among the variables is to use the Variance Inflation Factor (VIF) or to perform an OLS regression and examine the sign of the coefficients as shown in Table 3 given in the appendix. We found that the sign of OLS regression coefficients is opposite to that of the Pearson correlation. This mismatching in the OLS results are due to the high level of correlation between variables. This leads us to consider the PLS regression model to handle the problem of multicollinearity.

5 PLS Regression with Right-Censoring

In this section, we suggest to perform a PLS regression of $\tilde{Y} = (\tilde{y}_1, \ldots, \tilde{y}_n)$ describing a lifetime of firm until the failure on the matrix X composed of financial ratios. We just recall that the response variable can be right-censored. This is due to the fact that there are firms which have not been confronted to the failure event during the period of study but the event appeared afterwards. We apply the classical algorithm of the PLS regression of $\tilde{Y}$ on X for various censoring cases going from 5% to 70%. The results are interpreted in terms of model adjustment by evaluating the Fit Mean Squared Error (MSE_F), and in terms of the model predictive quality by estimating the Prediction Mean Squared Error (MSE_P). In the context of censored data, the MSE_F and the MSE_P are calculated using respectively the following algorithms:

Step 1: calculate $\hat{\tilde{y}}_i$ the ith observation PLS predicted value of $\tilde{y}_i$

Step 2: calculate $Z_i = \delta_i \left(\hat{\tilde{y}}_i - \tilde{y}_i\right)^2$

Step 3: calculate $Z = \sum_{i=1}^{n} Z_i$

Step 4: calculate $MSE_F = \frac{Z}{\sum_{i=1}^{n} \delta_i}$

and,

Step 1: calculate $\hat{\tilde{y}}_{i,-}$ the PLS regression predicted value after eliminating the observation i

Step 2: calculate $Z_i = \delta_i \left(\hat{\tilde{y}}_{i,-} - \tilde{y}_i\right)^2$

Step 3: calculate $\hat{S}^C(T_i^{C}-^{1})$ which is the Kaplan Meier estimator (1) calculated using the indicators $(1-\delta)$, where $\delta = \begin{cases} 1 \text{ if the observation is uncensored} \\ 0 \text{ otherwise} \end{cases}$

Step 4: calculate $Z = \sum_{i=1}^{n} \frac{\delta_i Z_i}{\hat{S}^C(T_i^{C}-)}$

Step 5: calculate $MSE_P = \frac{Z}{n}$

We provide below, in Figs. 1 and 2, respectively the values of MSE_F and MSE_P for various censoring levels, computed on the ten first PLS components.

Figure 1 shows that for data associated with a range of censoring rates going from 0 to 30%, the MSE_F decreases when increasing the number of components. This means that the adjustment quality of the PLS model improves when incorporating the components. Beyond a threshold censoring value of 30%, the associated MSE_F is going to decrease for a number of terms going from one to six.

Figure 2 shows that for one, two and three PLS terms, the uncensored data give biased prediction results. This means that considering censoring in the data

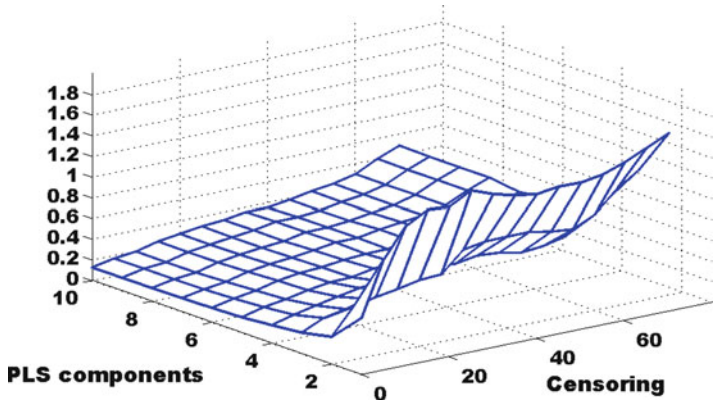


Fig. 1 MSE_F for the different censoring levels

[1] The symbol $(-)$ indicates the limit to the left of T_i^C.

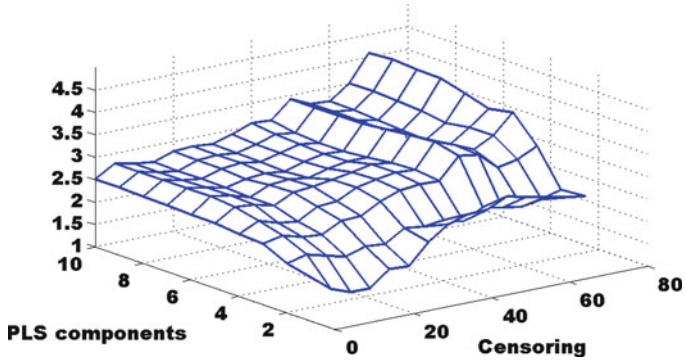


Fig. 2 MSE_P for the different censoring levels

improves the predictive quality of the model. In particular for one PLS component, the MSE_P with a range of censoring rates going from 5 to 25% is lower than that with 0% censoring rate. In the case of two components, the model is more successful for censoring rates of 15 and 25% than a model without censoring. For a model with three terms, it is more adequate to consider censoring rates of 10 and 15% as they give the lowest MSE_P. Finally, we observe that a model with 15% censoring rate and one PLS term seems to be the most adequate as it is associated with the minimal value of MSE_P. For this model, the estimated equation of the PLS regression is given by: $\hat{\hat{Y}} = 0.27\, t_1$, where the component t_1 describes the ratios to 28.17% and explains the lifetime of company until the failure event occurrence for 72.91%. It should be stressed here that, in contrast to the OLS coefficients, the signs of the PLS coefficients coincide with those of the Pearson correlation as shown in Table 3. Thus we can conclude that the PLS can cope with the multicollinearity problem. We can further improve the predictive quality of this model by calculating for every variable its explanatory power on the response variable. This is known as the Variable Importance in the Prediction (VIP) (Tenenhaus 1998).[2] According to the VIP criterion, we retain eleven ratios. Thus, the estimated equation becomes: $\hat{\hat{Y}} = 0.35\, t_1$ otherwise,

$$\hat{\hat{Y}} = 0.109x_9 + 0.102x_{10} + 0.083x_{11} + 0.115x_{15} + 0.122x_{16} + 0.127x_{17} + 0.098x_{22} + 0.096x_{28} + 0.090x_{30} + 0.082x_{31} + 0.128x_{34} \quad (2)$$

where the term t_1 describes the ratios up to 55.31% and explains the lifetime of firm until failure up to 75%. While interpreting the obtained results from (2), we see that the higher the levels of liquidity and cash, the lower the risk to go bankrupt and the longer the expected survival time. Besides, more the firms have an important equity,

[2] Variables having a VIP > 1 are the most important in the construction of the response variable. Therefore we can eliminate the variables with VIP values lower than 1.

more they will be reliable. Thus, they will have a higher longevity. In addition, we observe that the higher the performances in turnover, assets and added values, the lower the risk of distress.

In the sequel we intend to examine the prediction performance of the proposed model by predicting the lifetime of Tunisian Small and Medium-sized firms. The prediction is made using the PLS regression model with 15% censoring rate and one PLS term. Thus, the prediction set contains the remaining fifty-one observations. We have computed the *RMSE* for both OLS and PLS regression methods. The *RMSE* for the OLS-based method is 5.78 whereas the $RMSE$ of the PLS regression model is found to be markedly smaller with a value of 3.03. This confirms the superior prediction performance of the proposed method. In sum, the suggested model can be exploited to predict adequately the lifetime until failure for firms.

6 Conclusion

In this paper, we have discussed the problem of predicting lifetime of firms until failure using financial ratios as predictors within the PLS regression framework. We saw that problems may arise in practice due to right-censoring. So, the Mean imputation approach has been proposed to cope with these problems. We showed that taking into account the censoring in the data improves the predictive quality of the PLS model. Furthermore, the best model for predicting the lifetime until failure event is found to be a PLS model with a censoring rate of 15% and a single PLS term. We note in this paper that we have used the Mean imputation approach which estimates unobserved response variable from Kaplan Meier nonparametric estimator of the survival function. However, the Kaplan Meier estimator is an unstrict estimator which gives discontinuous estimation. In future work we intend to use another estimator to cope with this insufficiency.

Acknowledgements We would like to thank Dr. François-Xavier LEJEUNE for his valuable help.

Appendix

Table 3 VIF, OLS and PLS results[a]

Variables	VIF	Pearson correlation	OLS Parameters	PLS Parameters
x_7	5.3163	0.5942	−0.0168	0.0565
x_{10}	7.5799	0.4108	−0.0724	0.0391
x_{17}	5.3606	0.6120	−0.3510	0.0582
x_{29}	12.5473	−0.5465	1.0163	−0.0520

[a]We have reported some variables with VIF> 5 whose sign of OLS regression coefficients is opposite to that of the Pearson correlation

Table 4 Financial ratios

x_1	reserve/total assets	x_2	total liabilities/total assets
x_3	short-term debt/total debt	x_4	equity capital/total liabilities
x_5	gross operating surplus/total assets	x_6	net income/equity capital
x_7	net income/turnover	x_8	gross operating surplus/turnover
x_9	value added/turnover	x_{10}	current assets/short-term debt
x_{11}	cash/short-term debt	x_{12}	cash+marketable securities/short-term debt
x_{13}	cash flow/short-term debt	x_{14}	cash flow/total liabilities
x_{15}	cash/total assets	x_{16}	cash/current assets
x_{17}	cash/turnover	x_{18}	account receivable/total assets
x_{19}	working capital/total assets	x_{20}	account receivable+stock/total suppliers purchases
x_{21}	equity capital/permanent capital	x_{22}	equity/total assets
x_{23}	cash/bank loans	x_{24}	value added/tangible assets
x_{25}	financial charges/turnover	x_{26}	payroll/gross operating surplus
x_{27}	financial charges/gross operating surplus	x_{28}	turnover/total assets
x_{29}	non current assets/total assets	x_{30}	$turnover_t - turnover_{t-1}/turnover_{t-1}$
x_{31}	$totalassets_t - totalassets_{t-1}/totalassets_{t-1}$	x_{32}	$noncurrentassets_t - noncurrentassets_{t-1}/noncurrentassets_{t-1}$
x_{33}	$currentassets_t - currentassets_{t-1}/currentassets_{t-1}$	x_{34}	$valueadded_t - valueadded_{t-1}/valueadded_{t-1}$
x_{35}	Log(total assets)		

Source: These ratios are sourced from (Hamza and Baghdadi (2008)).

References

Altman, E. (1968). Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy. *Journal of Finance*, *23*, 589–609.

Bastien, P. (2008). Régression PLS et données censurées. Thése en informatique. Conservatoire National des Arts et Métiers, Paris.

Datta, S., Le-Rademacher, J., & Datta, S. (2007). Predicting patient survival from microarray data by accelerated failure time modeling using partial least square and LASSO. *Biometrics*, *63*, 259–271.

Hamza, T. & Baghdadi, K.(2008). Profil et déterminants financiére de la défaillance des PME Tunisiennes (1999–2003). *Banque et Marchés. Appl.93 (Mars-Avril)*, 45–62 .

Hillegeist, S., Cram, D., Keating, E., & Lundstedt, K. (2004). Assessing the probability of bankruptcy. *Review of Accounting Studies*, *9(1)*, 5–34.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Application to non orthogonal problems. *Technometrics*, *12*, 69–82.

Hougaard, P. (2000). *Analysis of multivariate survival data.* Springer-Verlag, New York-Berlin-Heidelberg.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observation. *Journal of American Statistical Association*, *53*, 457–81.

Massy, W.F. (1965). Principal components regression in exploratory statistical research. *Journal of American Statistical Association*, *60*, 234–256.

Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, *18*, 109–131.

Tenenhaus, M. (1998). *La régression PLS Théorie et pratique.* Paris: Technip.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of Royal Statistical Society*, *58*, 267–288.

Wold, H. (1966). *Estimation of principal component and related models by iterative least squares in multivariate analysis* (pp. 391–420). New York: Academic.

Robust Portfolio Asset Allocation

Luigi Grossi and Fabrizio Laurini

Abstract Selection of stocks in a portfolio of shares represents a very interesting problem of 'optimal classification'. Often such optimal allocation is determined by second-order conditions which are very sensitive to outliers. Classical Markowitz estimators of the covariance matrix seem to provide poor results in financial management, so we propose an alternative way of weighting observations by using a forward search approach. An application to real data, which shows the advantages of the proposed approach is given at the end of this work.

1 Introduction

The Markowitz mean-variance efficient frontier is the standard theoretical model for normative investment behavior Markowitz (1959). In this paper, we discuss the problem of statistical robustness of the Markowitz optimizer and show that the latter is not robust, meaning that a few extreme assets prices or returns can lead to irrelevant 'optimal' portfolios. We then propose a robust Markowitz optimizer and show that it is far more stable than the classical version based on Maximum Likelihood Estimator (MLE). Suppose there are N risky assets whose prices observed for T periods are $p_{it}, t = 1, \ldots, T, i = 1, \ldots, N$ and let $\mathbf{x} = (x_1, \ldots, x_N)'$ be the vector of portfolio weights. The assets returns are given by a matrix $\mathbf{Y} = (\mathbf{y}_t^{(1)}, \ldots, \mathbf{y}_t^{(N)})'$, where $\mathbf{y}_t^{(i)} = (y_{i1}, \ldots, y_{iT})'$ and $y_{it} = log(p_{it}/p_{it-1})$ with expected returns given by a $N \times 1$ vector $\boldsymbol{\mu}$ and $N \times N$ covariance matrix $\boldsymbol{\Sigma}$. The expected return and variance of the portfolio can be written as $\mu_p = \mathbf{x}'\boldsymbol{\mu}$ and $\sigma_p^2 = \mathbf{x}'\boldsymbol{\Sigma}\mathbf{x}$, respectively.

For a given level of risk-aversion γ, the classical mean-variance optimization problem can be formulated as $\max_{\boldsymbol{x}}(\mu_p - \gamma\sigma_p^2)$, subject to the constraints $\mathbf{x} \geq 0$ and $\mathbf{x}'\boldsymbol{\iota}_N = 1$, where $\boldsymbol{\iota}_N$ is a $N \times 1$ vector of 1. The constraint of no short-selling ($\mathbf{x} \geq 0$) is very frequently imposed as many funds and institutional investors are not allowed to sell stocks short.

The expected returns, variances and covariances are usually estimated by means of the classical maximum likelihood estimators.

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_34,

One problem with the standard MLEs is that they are sensitive to influential observations, typically represented by extreme returns. In this paper we suggest to use a robust weighted version of the covariance matrix estimator where weights are obtained through a forward search procedure. Other papers have suggested portfolio optimization based on robust estimators. See Fabozzi et al. (2007), Welsch and Zhou (2007) and DeMiguel and Nogales (2009). Their approaches differ from ours because they apply robust estimators, such as winsorized, M- and S- estimators, Minimum Volume Ellipsoids which have a high breakdown point, but are not as efficient as MLEs when the underlying model is correct. The forward search, instead, solves this problem as it combines the efficiency of MLEs to the outlier resistance of very robust estimators.

2 Robust Weighted Covariance Matrix

Our target is to compute weights $w_t \in [0, 1]$, for each observation in the multiple time series $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{Nt})'$, $t = 1, \ldots, T$, with the forward search method (Atkinson and Riani, 2000; Atkinson et al., 2004). The weights will be used to obtain a weighted covariance matrix such that the most outlying observations get small weight. For multivariate data, standard methods for outlier detection are based on the squared Mahalanobis distance for the t-th observation: $d_t^2 = (\mathbf{y}_t - \hat{\boldsymbol{\mu}})' \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{y}_t - \hat{\boldsymbol{\mu}})$, where both the mean and variance are estimated. The goal of the forward search is the detection of units which are different from the main bulk of the observations and to assess the effect of these units on inferences made about the model. Some methods for the detection of multiple outliers, like the forward search applied in this paper, use very robust methods to sort the data into a clean part (CDS = Clean Data Set) of size m_0 and potential outliers of size $T - m_0$. Given the best subset $S_*^{(m)}$ of dimension $m \geq m_0$ detected at step m, we can calculate a set of T squared Mahalanobis distances, defined as

$$d_{t(m)}^2 = (\mathbf{y}_t - \hat{\boldsymbol{\mu}}_m)' (\hat{\boldsymbol{\Sigma}}_m)^{-1} (\mathbf{y}_t - \hat{\boldsymbol{\mu}}_m) \tag{1}$$

where $\hat{\boldsymbol{\mu}}_m$ and $\hat{\boldsymbol{\Sigma}}_m$ are the mean and covariance matrix estimated on the m-sized subset. We then increase the size of the initial CDS selecting observations with small Mahalanobis distances and so are unlikely to be outliers. Thus, with the forward search algorithm the data are ordered according to their degree of agreement with the underlying model, with observations furthest from it joining the CDS in the last steps of the procedure. It is common, during the forward search to graphically monitor $d_{t(m)}^2$ as m increases from m_0 to T; sharp changes of the curves indicate the introduction of influential observations into the CDS. When $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are MLE on the whole sample, the classical Mahalanobis distances follow a scaled beta distribution. But in (1) the Mahalanobis distances are estimated from a subset of m observations which do not include the observation being tested. In such a case, the reference null distribution would be (see, Atkinson et al. 2009):

$d^2_{t(m)} \sim [T/(T-1)][N(m-1)/(m-N)]F_{N,T-N}$, where N is the number of columns of $\boldsymbol{Y}$. The weights are computed for each observation at the end of the forward search. For the computation of the weights we compare the trajectories of $d^2_{t(m)}$ during the forward search with confidence bands from the F distribution. Alternatively, simulated envelopes can be adopted. At each step of the forward search, we measure the degree of outlyingness of each observation $t = 1, \ldots, T$, as the squared Euclidean distance, π, between the distance (1) lying outside a confidence band and the boundaries of the band itself by considering the F distribution with $N, T-N$ degrees of freedom, and the quantile F_δ at the nominal level $1-\delta$. For a fixed step of the forward search m, we record the distance of the t-th trajectory from the percentile of the confidence band, provided that the t-th observation is outside the $1-\delta$ nominal level. If $d^2_{t(m)}$ lies under the F_δ percentile, then, at the m-th step, it will get zero distance. At the next step $m+1$, the weight of the t-th observation will be increased by an amount which is induced by the squared Euclidean distance between the t-th trajectory and the percentile of the confidence band, provided that at step $m+1$ the t-th trajectory exceeds the nominal level $1-\delta$. If at step $(m+1)$-th the t-th trajectory lies under the F_δ quantile, then a zero will be added to the distance computed at the step m.

The overall degree of outlyingness for the t-th observation is given by the sum of all squared Euclidean distances, computed only when the trajectory exceeds the confidence bands. Formally, letting $\pi_m^{(t)}$ be the distance between $d^2_{t(m)}$ and the percentile F_δ, for the unit t-th at the step m, we define the squared Euclidean distance as: $\pi_m^{(t)} = 0$ if $d^2_{t(m)} \in [0, F_\delta]$, $\pi_m^{(t)} = (d^2_{t(m)} - F_\delta)^2$ if $d^2_{t(m)} > F_\delta$ and we consider the overall distance of the t-th observation as the sum of such distances during the forward search, i.e.,

$$\pi_t = \frac{\sum_{m=m_0}^{T} \pi_m^{(t)}}{T - m_0 + 1}. \tag{2}$$

The squared Euclidean distance measures the degree of outlyingness of each observation through the computation of a weight, in the interval $[0, 1]$, obtained with the following mapping of (2): $w_t = \exp(-\pi_t)$. The last step of the procedure is based on the creation of a weighted covariance matrix to be used in the optimization. For a given series of returns i, with $i = 1, \ldots, N$, we consider the weighted return $y^\star_{it} = y_{it} w_t^{1/2}$. The weighted covariance matrix will be simply the covariance matrix based on the weighted observations, that we denote as $\tilde{W}$. Notice that also the sample mean will be modified by using weighted returns.

3 Monte Carlo Experiment

In this section we are going to introduce some results of a Monte Carlo experiment which has been carried out to highlight the main advantages which come from the application of the forward search based estimator (FWD from now on) with respect to the classical non robust MLE. Let us suppose there are six securities (that is

$N = 6$) with true parameters, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, chosen to be in average range of monthly returns of many securities. A monthly frequency was chosen as a reasonable intermediate point between weekly and yearly intervals. A weekly time horizon is too short for most portfolio planning problems and an annual time horizon is too long for gathering historical data. However the same analysis can be applied to any chosen time horizon. Then $T = 120$ monthly returns for each security have been generated. In order to put in evidence the effect of outliers we contaminated the series with additive outliers at random positions according to the following equation:

$$\boldsymbol{y}_t^* = \boldsymbol{y}_t + \theta \boldsymbol{\delta}_t$$

where $\boldsymbol{y}_t \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\delta}_t$ is a stochastic contamination process, which takes non-zero values with positive probability, and where θ is a non-zero constant indicating the magnitude of the additive outliers.

One of the most interesting output of mean-variance portfolio analysis is the frontier where optimal combinations of returns and risk levels are reported for given levels of risk aversion. In simulation experiments it is possible to compare several versions of efficient frontiers (see, Broadie, 1993) which are reported as follows:

- True efficient frontier
$$\mu_p = \boldsymbol{x}'\boldsymbol{\mu} \qquad \sigma_p^2 = \boldsymbol{x}'\boldsymbol{\Sigma}\boldsymbol{x} \tag{3}$$
- Estimated frontier
$$\mu_p = \hat{\boldsymbol{x}}'\hat{\boldsymbol{\mu}} \qquad \sigma_p^2 = \hat{\boldsymbol{x}}'\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{x}} \tag{4}$$
- Actual frontier
$$\mu_p = \hat{\boldsymbol{x}}'\boldsymbol{\mu} \qquad \sigma_p^2 = \hat{\boldsymbol{x}}'\boldsymbol{\Sigma}\hat{\boldsymbol{x}} \tag{5}$$

The true frontier in (3) is not observable with real data. Next, using the estimated parameters, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$, the estimated frontier (4) is obtained. Finally, the actual frontier (5) is computed combining weights from the estimated frontier and true parameters. To summarize, the true frontier is the efficient frontier based on the true but unknown parameters, the estimated is the frontier based on the estimated and hence incorrect parameters, the actual frontier consists of the true portfolio means and variance points corresponding to portfolios on the estimated frontier. In short, the estimated frontier is what appears to be the case based on the data and the estimated parameters, but the actual frontier is what really occurs based on the true parameters. The estimated frontier is the only frontier that is observable in practice.

Figure 1 reports an example of the three frontiers defined earlier based on simulated uncontaminated data and estimated with MLE. What it is worth to be noticed is that limiting our attention to optimal portfolios (between two portfolios with equal mean, the less risky is chosen), the estimated frontier always lies above the true frontier. This illustrates the error maximization property of mean variance analysis. That is, points on the estimated frontier correspond to portfolios that overweight securities with large positive mean returns, large negative errors in standard deviations and large negative errors in correlations. These estimation errors lead to optimistically biased estimates of portfolio performance. On the other hand, the actual frontier

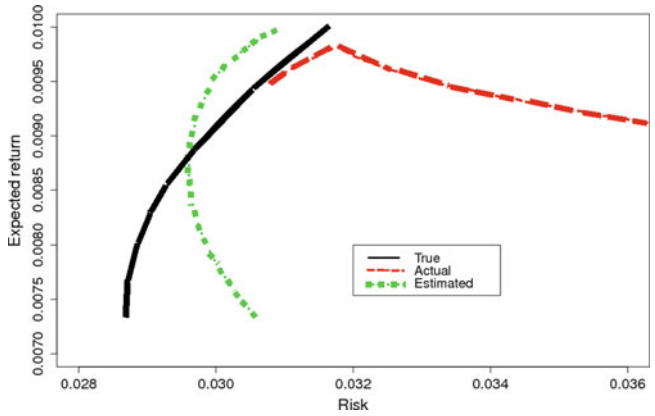


Fig. 1 True, estimated and actual frontier on a simulated data set

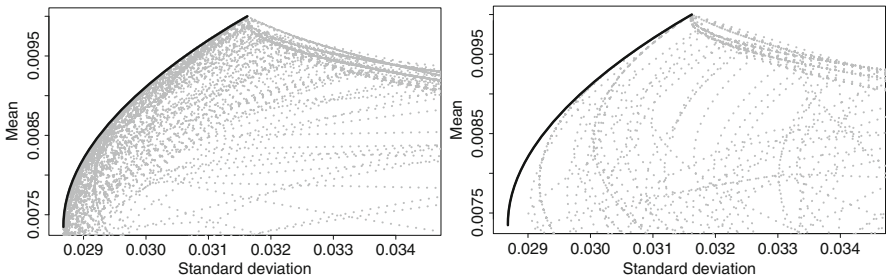


Fig. 2 True and actual MLE frontiers for uncontaminated (*left plot*) and contaminated (*right panel*) data. The true frontier is drawn by a solid black line, dotted lines are used for the actual frontiers

lies under the true frontier. This example shows that generally minimum variance portfolios can be estimated more accurately than maximum return portfolios.

What has been shown in Fig. 1 is simply based on one simulation. Thus, we repeated the simulation experiment (1,000 trials) generating random monthly returns based on the same true parameters, estimating model parameters from the simulated returns and then computing actual mean-variance frontiers.

In Fig. 2 actual frontiers (dashed lines) for a fraction of the total trials are compared with the true frontier (solid line). The left panel has been obtained simulating uncontaminated time series of returns, while the right panel refers to the case where simulated time series have been contaminated with outliers. The effect of contamination consists of spreading out the actual frontiers far from the true frontier. This gives an idea about the impact of outliers on MLE.

When more than one trial is carried out we have to summarize the estimation error. To this aim we can focus on individual target points on the true efficient frontier. The target point is defined as the point in the mean-variance plane that

maximizes $\mu_p - \gamma\sigma_p$ for a given value of risk aversion γ. For $\gamma = 0$ the solution is the portfolio with the maximum return. When $\gamma \to \infty$ the solution is the minimum variance portfolio. In order to a have a quantitative measure of the error caused by using estimated parameters, we compute the distance between the target point on the true frontier and the corresponding point on the actual. Finally, we obtain the following Root Mean Square Error (RMSE):

$$\Delta_\mu(\gamma) = \sqrt{\frac{1}{S}\sum_{s=1}^{S}[\mu_p(\gamma) - \tilde{\mu}_p^s(\gamma)]^2} \quad \Rightarrow RMSE \quad \text{for} \quad \mu_p$$

$$\Delta_\sigma(\gamma) = \sqrt{\frac{1}{S}\sum_{s=1}^{S}[\sigma_p(\gamma) - \tilde{\sigma}_p^s(\gamma)]^2} \quad \Rightarrow RMSE \quad \text{for} \quad \sigma_p$$

where S = number of simulations, $(\mu_p(\gamma)\mu_p(\gamma), \sigma_p(\gamma)\sigma_p(\gamma))$ is the target point on the true frontier and $(\tilde{\mu}_p^s(\gamma), \tilde{\sigma}_p^s(\gamma))$ is the target point on the actual frontier.

In Fig. 3 the evolution of the RMSE with 1,000 simulations of μ and σ for MLE and forward search estimators (FWD) are reported for uncontaminated data as function of $\log(\gamma)$. The continuous line is for MLE, while the dashed line is for FWD. As can be noticed, when time series are free from outlier MLE and FWD lead to very similar results in terms of RMSE with a slight prevalence of MLE on FWD in the case of σ.

Finally, Fig. 4 reports the evolution of RMSE when MLE, and FWD for mean-variance parameters have been applied. These plots are similar to those shown in Fig. 3 but refer to contaminated data where patches of consecutive outliers have been placed in each security time series at random positions. The left plot reports RMSE for μ while the right plot reports RMSE for σ. Solid line is for MLE and dashed for FWD. It is extremely clear that, in the case of contaminated data, the FWD estimator

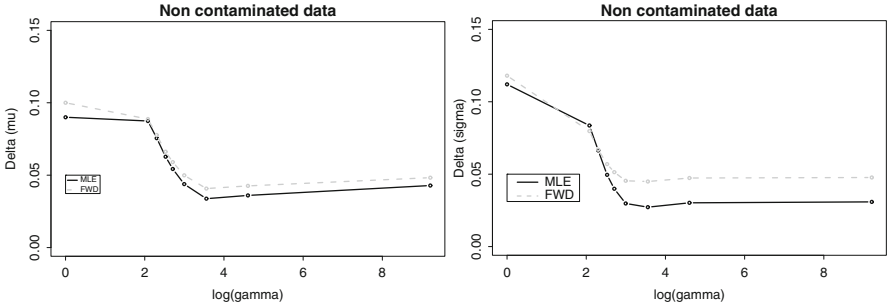


Fig. 3 RMSE for μ (*left panel*) and σ (*right panel*) estimated with MLE and forward search on uncontaminated data

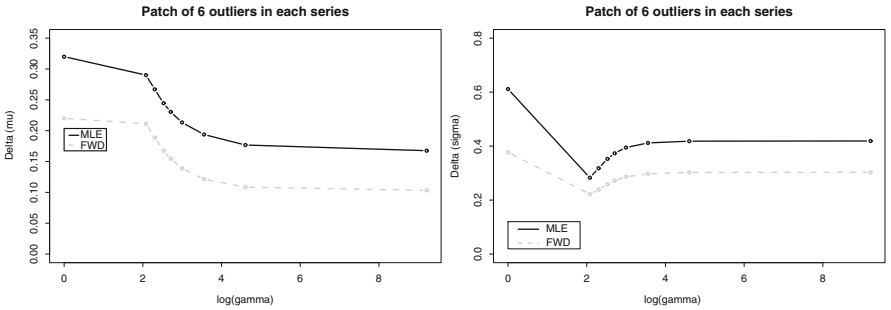


Fig. 4 RMSE (μ, *left panel*, and σ, *right panel*) for MLE and FWD when data are contaminated with patches of outliers

leads always to the lowest estimation error and this should be reflected in a better portfolio allocation.

4 Application to Financial Data

We consider the monthly returns of six stocks of the US market with data from January 1973 to March 2009 included. Data come from Datastream. We have computed the efficient frontier according to the so called 'tangency' optimality, i.e., using the Sharpe ratio for the optimal weights allocation of each asset into the final portfolio. The efficient frontier is computed by standard MLE and by the forward weighted estimation of the covariance matrix. A very remarkable feature of our approach is that, when anomalous observations are weighted with our method, the overall 'portfolio risk', measured by the standard deviation, is massively reduced compared with other methods. However, even if we have a much lower risk portfolio the expected returns of the frontier is still broadly consistent with the classical method. In conclusion, with our forward method of weighting observations we have a preferable portfolio, which has the same expected return with a much lower risk. In order to show the importance of using robust estimators in portfolio allocation we compare the performance of allocations derived from the analyzed estimators and using a rolling windows technique explained as follows: (1) estimate weights (MLE and FWD) using data for $t = 1, \ldots, T-1$ and get the average portfolio return in T. (2) estimate weights (MLE and FWD) using data for $t = 2, \ldots, T$ and get the average portfolio return in $T+1$. (3) estimate weights (MLE and FWD) using data for $t = d, \ldots, T-d$ and get the average portfolio return in $T-d+1$.

The right panel of Fig. 5 reports the output of the rolling windows procedure estimating parameters on data until the end of 2005 and taking year 2006 as a forecast period according to the rolling windows procedure. The dashed line is for FWD while solid is for MLE. Portfolio performances are generally better when the forward search weights are applied.

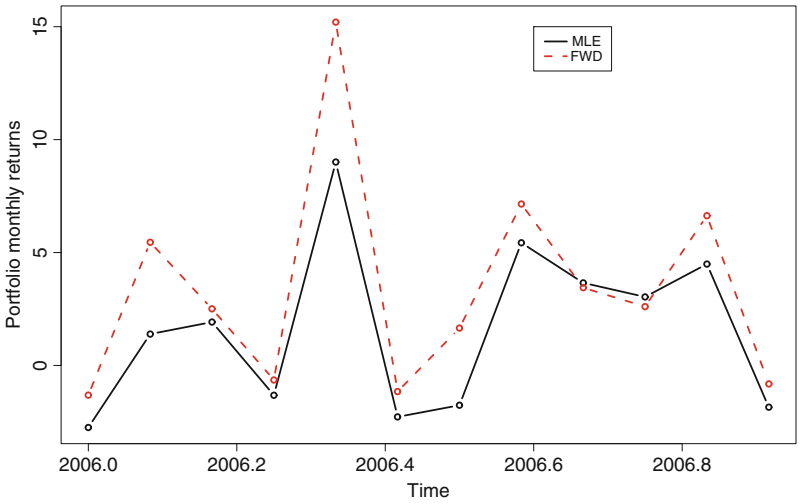


Fig. 5 Portfolio monthly performances in 2006 using a rolling windows technique

5 Conclusions and Further Research

In this paper a new robust approach to the asset allocation problem has been suggested. Simulation experiments lead to very promising results. Further experiments are needed to better understand how the robust forward search estimator can improve the performance of portfolio allocation, but the possibility to compare true, actual and estimated frontiers is a very easy way. On the other hand, we have to devote more time in analyzing real data and in developing new measures for assessing the superiority of robust methods in solving allocation problems because only estimated frontiers are observable. This implies that the forecasting properties of robust methods should be analyzed. When back-tested on market data, these methods are shown to be effective in improving portfolio performance. Robust asset allocation methods have great potential to improve risk-adjusted portfolio returns and therefore deserve further exploration in investment management research.

References

Atkinson, A.C. & Riani, M., (2000). *Robust Diagnostic Regression Analysis*. New York: Springer-Verlag.

Atkinson, A. C., Riani, M., & Cerioli, A. (2004). *Exploring multivariate data with the forward search*. New York: Springer–Verlag.

Atkinson, A. C., Riani, M., & Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B*, *71*(2), 447–466.

Broadie, M. (1993). Computing efficient frontiers using estimated parameters. *Annals of Operations Research*, *45*(1), 21–58.

DeMiguel, V., & Nogales, F. J. (2009). Portfolio selection with robust estimation. *Operations Research*, *57*(3), 560–577.

Fabozzi, F. J., Kolm, P. N., Pachamanova, D. A., & Focardi, S. M. (2007). *Robust Portfolio Optimization Management*. New York: Wiley.

Markowitz, H. (1959). *Portfolio selection: Efficient diversification of investments*. New York: Wiley.

Welsch, R. E., & Zhou, X. (2007). Application of robust statistics to asset allocation models. *REVSTAT: Statistical Journal*, *5*(1), 97–114, http://www.ine.pt/revstat/pdf/rs070106.pdf

A Dynamic Analysis of Stock Markets through Multivariate Latent Markov Models

Michele Costa and Luca De Angelis

Abstract A correct classification of financial products represents the essential and required step for achieving optimal investment decisions. The first goal in portfolio analysis should be the allocation of each asset into a class which groups investment opportunities characterized by a homogenous risk-return profile. Furthermore, the second goal should be the assessment of the stability of the classes composition. In this paper we address both objectives by means of the latent Markov models, which allow us to investigate the dynamic pattern of financial time series through an innovative framework. First, we propose to exploit the potential of latent Markov models in order to achieve latent classes able to group stocks with a similar risk-return profiles. Second, we interpret the transition probabilities estimated within latent Markov models as the probabilities of switching between the well-known states of financial markets: the upward trend, the downward trend and the lateral phases. Our results allow us both to discriminate the stock's performance following a powerful classification approach and to assess the stock's dynamics by predicting which state is going to experience next.

1 Introduction

Modern portfolio theory suggests to evaluate financial products on the basis of two latent variables: the risk and the expected return. Frequently unexpected, and sometimes disastrous, outcomes of investment decisions derive from the latent nature of this information set, or, more correctly, from the difficulty to provide a satisfactory measurement of the risk and the expected return. In order to achieve this goal, in this paper we propose, as a preliminary and required step for investment decisions, to classify financial products into latent classes characterized by a homogenous risk-expected return profile. We focus on the latent Markov models, which allow us to investigate the dynamic pattern of financial time series (Dias et al. 2009, Rydén et al. 1998). With respect to mainstream literature on Markov switching models (e.g., Hamilton and Susmel 1994; Rossi and Gallo 2006), our proposal allows a specific focus on the latent features of the different stock market phases using a

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_35,

simpler model specification, in which we do not specify a particular model such as AR or ARCH-type. Therefore, LMM enables an important reduction of the number of parameters to be estimated which facilitates the exploratory investigation of the latent stochastic process underlying the observed time series.

In this framework, we are able to

1. Propose an innovative measurement of risk and expected return,
2. Differentiate risk and expected return evaluation with respect to stock market phases,
3. Provide transition probabilities between different stock market phases.

Furthermore, combining the information provided by each latent state's profile and the latent transition probabilities, we can predict quite accurately which state the stock market is going to experience in the future, thus enabling new opportunities in investment decisions. Finally, our proposal allows us to advance the discussion about the role of statistical methodology in the analysis of financial variables.

2 The Latent Markov Model

The Latent Markov Model (LMM), also known as Hidden Markov Model (Baum et al. 1970) or regime-switching model (Hamilton and Raj 2002), is a powerful tool for describing the dynamics of a set of observed responses. Denoting by Z_t the vector containing the return observations of n stock market indexes at time t (for $t = 1, \ldots, T$), where the vector element z_{it} denotes the return observation of index i (for $i = 1, \ldots, n$) at time t, the LMM analyzes $f(Z)$, the probability density function of the return distribution of vector Z over time, by means of a latent transition structure defined by a first-order Markov process. For each time point t, the model defines one discrete latent variable denoted by Y_t constituted by S latent classes (which are referred to as latent states), thus overall the LMM includes T latent variables.

The LMM is specified as

$$f(Z) = \sum_{Y_1=1}^{S} \cdots \sum_{Y_T=1}^{S} f(Y_1) \prod_{t=2}^{T} f(Y_t|Y_{t-1}) \prod_{t=1}^{T} f(Z_t|Y_t) \tag{1}$$

The model in (1) relies on two main assumptions: first, it assumes that the sequence of the latent states Y_t for $t = 1, \ldots, T$ follows a first-order Markov chain, i.e., Y_t is associated only with Y_{t-1} and Y_{t+1}; second, the observation at a particular time point is independent of observations at other time points conditionally on the latent state Y_t. The latter implies that the observed index return at time t depends only on the latent state at time t and it is often referred to as the local independence assumption which is the pillar of latent structure models.

The LMM is characterized by three probability functions:

1. $f(Y_1)$ is the (latent) initial-state probability;
2. $f(Y_t|Y_{t-1})$ is a latent transition probability which denotes the probability of being in a particular latent state at time t conditional on the state at time $t-1$. Assuming a homogenous transition process with respect to time, we achieve the latent transition matrix P where the generic element p_{jk}, with $j,k = 1,\ldots,S$, denotes the probability of switching from latent state j to latent state k.
3. $f(z_{it}|Y_t)$ is the Gaussian density function for the observation, that is the probability density of having a particular observed return of index i at time t conditional on the latent state occupied at time t. This distribution is characterized by a parameter vector $\theta_j = (\mu_j, \sigma_j)$ which contains the mean and the standard deviation for the latent state j and which highlights the feature of each latent state.
   The multiplication of the T density functions in (1) denotes the measurement component of the model.

The model with such specification allows us to discriminate different latent states which define an underlying unobserved stochastic process common to the n indexes achieved on the basis of the observed return distributions of the indexes. Therefore, each latent state can be considered as a particular phase of the entire stock market. We want to stress the multivariate specification of the model which takes into account not a single observed variable but the $n \times 1$ vector Z. In this multivariate framework, the LMM analyzes the covariance structure existing in the observed data providing the estimation of a latent stochastic process defined by Y_t which accounts for the common dynamic pattern of the n observed variables.

The number of free parameters ($NPar$) of the multivariate LMM is $S(S+2n)-1$, that is $S-1$ initial-states, $S(S-1)$ transition probabilities, and $2nS$ conditional means and standard deviations of the observed variables.

The LMM is estimated by means of a variant of the EM procedure, the forward-backward or Baum-Welch algorithm (Baum et al. 1970) which exploits the conditional independencies implied by the model in order to circumvent the computational problem due to high values of T. This variant was extended by Paas et al. (2007) in order to deal with multiple observed indicators. This extension is implemented in the Latent GOLD 4.5 computer program Vermunt and Magidson (2007).

3 Model Estimation and Results

We apply the LMM described above to a data set concerning the monthly return distribution from February 2002 to December 2008 for a total of $T = 83$ observations of $n = 5$ benchmark Italian indexes.

Each index represents a different stock market segment characterized by a particular level of capitalization: in decreasing order, we consider S&PMIB, MIDEX,

Table 1 Estimation results from multivariate LMMs with different number of latent states

Latent States	*LL*	*NPar*	*CAIC*
6	−913.89	95	1922.79
7	−887.81	118	1893.63
8	−862.63	143	1868.27
9	−836.30	170	1842.60
10	−826.57	199	1852.14
11	−810.58	230	1851.16

ALLSTARS, and EXPANDI. We also analyze the MIBTEL index which includes all the stocks traded on the Italian stock market.

We fit a conditional independent LMM where the five stock indexes are independent conditionally on the latent states. The model with such specification allows us to define the unobserved stochastic process common to the five market indexes and composed by S different latent states in which are classified the time points characterized by similar return observations of the indexes. The measurement of S represents quite an important issue and a relevant result, because, until now, there is no answer to the simple question of how many phases characterize the financial markets.

According to the Consistent AIC statistics (CAIC; Bozdogan 1987) reported in Table 1, the model which provide the best fit to the data is the LMM with $S = 9$ latent states ($LL = -836.3$, $NPar = 170$, $CAIC = 1842.6$). Hence, the model identifies $S = 9$ different market phases. This is a relevant number and it is likely we would have preferred a lesser number of states. However, we want to stress how our completely subjective opinion about the stock market regimes could be too optimistic and it is also possible that data generating processes in stock markets are not so simple as we would like. Moreover, during the period 2002–2008, stock markets have experienced many shocks and big crises and, therefore, nine phases could not be our preferred choice but lead to a consistent representation of the financial variable dynamics. Furthermore, the use of the CAIC statistic may rise some issues about its robustness. However, our framework introduces a rigorous methodological approach for defining the number of market regimes which allows us to avoid a subjective a priori decision.

Each latent state can be characterized by the mean return (μ_j) and standard deviation (σ_j) of the market indexes. The standard deviations provide information about the volatility (i.e., the risk) of each latent state: a low volatility state can be interpreted as a 'stable' market regime, while a state with a high standard deviation indicates a period of market turbulence.

In Table 2 the latent states are ranked according to the mean return. The first column in Table 2 provides the size of each latent state which indicates the proportion of time observations classified into a particular state and, thus, represents their level of occurrence in the period that is analyzed. For example, the 20% of the time points are allocated into state 8 which represents the modal state, while the state 9 contains only 4% of the observations.

Table 2 Measurement component of the multivariate LMM with nine latent states

Index	Size	S&PMIB		MIDEX		ALLSTARS		EXPANDI		MIBTEL		Average	
State		$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$
1	0.078	−12.80	2.13	−13.86	2.19	−10.68	1.90	−8.50	3.61	−12.29	1.31	−11.63	*2.23*
2	0.098	−6.68	3.36	−8.34	1.71	−6.35	2.56	−2.85	2.17	−6.71	2.17	−6.18	*2.39*
3	0.171	−3.39	1.09	−2.46	1.92	−1.45	1.85	−0.04	2.62	−2.94	0.93	−2.05	*1.68*
4	0.044	−0.20	0.99	0.37	3.40	−3.52	1.31	−1.41	3.77	−0.25	0.97	−1.00	*2.09*
5	0.116	−1.02	1.03	0.01	1.20	0.11	1.04	0.63	2.26	−0.64	0.99	−0.18	*1.30*
6	0.142	2.14	1.10	0.58	1.09	1.38	1.97	−0.04	0.83	1.75	1.09	*1.16*	*1.22*
7	0.106	4.82	0.90	4.33	2.75	3.40	2.34	0.79	1.78	4.56	0.83	*3.58*	*1.72*
8	0.204	2.84	0.85	5.20	1.66	4.60	1.69	3.41	2.50	3.21	0.77	*3.85*	*1.50*
9	0.043	9.76	1.53	8.17	3.13	5.60	1.70	1.21	0.96	8.72	1.34	*6.69*	*1.73*

The analysis of $(\hat{\mu}_j, \hat{\sigma}_j)$ allows powerful insight about the main features of the different phases which characterize the dynamics of the five stock market indexes. States 1 and 2, for example, identify financial crisis periods since they are characterized by a low value of $\hat{\mu}$ and a high value of $\hat{\sigma}$. However, according to their sizes, these two latent states do not occur often (8 and 10%, respectively). On the other hand, states 7, 8, and 9, with a high value of $\hat{\mu}$, refer to positive regimes. Latent states 5 and 6 are characterized by moderate values of $\hat{\mu}$ and the lowest values of $\hat{\sigma}$. These states represent the phases of stability of the stock market and together represent more than one fourth of the observations. Furthermore, from Table 2 it can be noted that, within each latent state, $\hat{\mu}$ and $\hat{\sigma}$ differ among the market indexes. In particular, the ALLSTARS and EXPANDI, which comprise the small capitalization companies, usually have lower values with respect to S&PMIB, MIDEX, and MIBTEL. The only exception is represented by latent state 4: this state groups only 4% of the observations and is characterized by strong negative values of $\hat{\mu}$ for the ALLSTARS and EXPANDI and values close to zero for the other indexes. Moreover, the values of $\hat{\sigma}_4$ for the MIDEX and EXPANDI are the highest ones. Hence, the LMM classifies into the latent state 4 the time points in which the index returns exhibit contrasting behaviour. Figure 1 displays actual and estimated time series by referring to the S&PMIB index and the LMM with nine latent states. The estimated series is plotted using the state return means for the S&PMIB. Figure 1 shows that the LMM approximates the index dynamic pattern quite accurately.

The results reported in Table 2 provide an innovative measurement of the risk and expected return and allow us to identify alternative portfolio investment strategies implied by the use of the different stock market indexes.

A further relevant set of information provided by the LMM is represented by the latent transition matrix P which shows the probability of switching from one latent state to another. The results related to the dynamics of the five Italian stock indexes are reported in Table 3. The values on the main diagonal of matrix P represent the state persistence, that is the probabilities of remaining in a particular market phase. For example, the probability of staying in latent state 2 is $p_{22} = .250$, while it is very unlikely to remain in state 3 ($p_{33} = .001$). The out of the diagonal p_{jk} values indicate the probabilities of market regime switching: for instance, when the stock

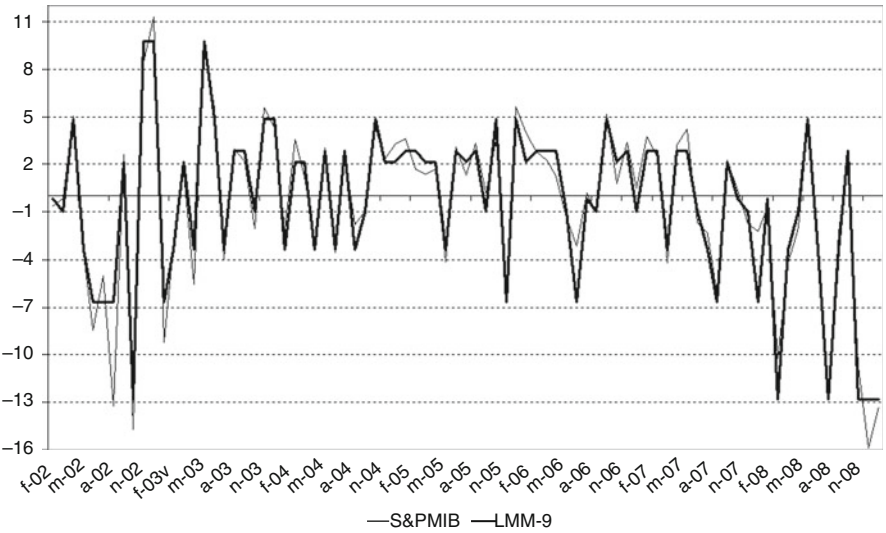


Fig. 1 S&PMIB and LMM estimate time series

Table 3 Latent transition matrix P

j/k	1	2	3	4	5	6	7	8	9
1	0.394	0.002	0.394	0.002	0.002	0.002	0.002	0.002	0.198
2	0.002	0.250	0.123	0.247	0.002	0.249	0.125	0.002	0.002
3	0.072	0.143	0.001	0.001	0.142	0.143	0.001	0.427	0.072
4	0.245	0.003	0.003	0.003	0.734	0.003	0.003	0.003	0.003
5	0.001	0.200	0.099	0.001	0.002	0.001	0.584	0.110	0.001
6	0.084	0.001	0.250	0.084	0.001	0.245	0.001	0.333	0.001
7	0.001	0.112	0.446	0.001	0.001	0.323	0.112	0.002	0.001
8	0.059	0.001	0.173	0.001	0.290	0.123	0.001	0.351	0.001
9	0.004	0.325	0.004	0.004	0.004	0.004	0.325	0.004	0.325

market is experiencing a crisis, represented by latent state 1, at time $t + 1$ it is quite likely a persistence of the negative market phase ($p_{11} = .394$ and $p_{13} = .394$) but also a switch to a strong positive regime ($p_{19} = .198$). It is also interesting to notice that when we are in the very positive market regime at time t it may persist also at time $t + 1$ with probability $p_{99} = .325$ and, with the same probability, it may also shift to a crisis period represented by state 2 or continue the positive phase switching to the state 7.

We can exploit the information provided by the switching probabilities reported in Table 3 in order to evaluate the reliability of the prediction capability of the LMM. Table 3 shows that some regime switching can be predicted quite accurately because their transition probabilities are high. For instance, the shift from latent state 4 to state 5 is relative easy to predict ($p_{45} = .734$). On the contrary, there are latent states for which at least four transition probabilities are above 0.05, which complicates

Table 4 Number of times and percentages LMM correctly predict the latent state at time $t + 1$ according to the five highest transition probabilities

Modal p_{jk}	I	II	III	IV	V	Others	Total
Count	34	21	14	8	4	1	82
Percentage	41.46	25.61	17.07	9.76	4.88	1.22	100

prediction. For example, latent state 3 has five transition probabilities higher than 0.05 and it makes market regime at time $t + 1$ quite difficult to predict precisely.

Following the approach provided by De Angelis and Paas (2009), since each regime switching has its own probability to occur, we can determine the LMM prediction power by referring to one-step ahead forecasts summarized in Table 4. In this table, we report the number of times the LMM is able to predict the next regime correctly according to the five highest latent transition probabilities. Thus, column I reports the amount of the times that LMM predicts the next market phase by referring to the most probable p_{jk} in matrix P, column II contains the amount of the times LMM forecasts correctly according to the second modal transition probability, and so on. For instance, August 2008 observation has been classified into latent state 3 and September 2008 into state 8. Since $p_{38} = .427$ is the highest probability for latent state 3, we reported this case in column I of Table 4. The second-last column of Table 4 provides the number of times that the model is unable to predict the next month regime by referring to the five most probable latent transition probabilities. It must be noted that the percentage of column 'Others' which can be considered as the proportion of times that, in a certain sense, LMM fails to predict the next market regime is only 1.22%. On the contrary, the model prediction accuracy based on the percentages of columns I, II, and III jointly almost reaches 85%.

4 Conclusions

The multivariate LMM defines different market regimes for the Italian stock market and provides the probabilities of switching between one regime and another. The knowledge of the latent state features and the transition probabilities is decisive in order to properly measure the latent risk-return profile of financial products. First, our approach provides a methodologically correct solution able to differentiate the latent evaluation with respect to the stock market phases. Second, within the LMM framework, it is also possible to analyze the dynamic pattern of the unobservable risk-return profile. The latent transition probabilities enable us to predict the market regime at time $t + 1$ quite accurately by referring to the highest values of p_{jk}. Furthermore, the classification of every time point of the time series in homogenous non-observable states offers a contribution in model-based clustering for financial time series (Frühwirth-Schnatter and Kaufmann 2008, Otranto 2008) which is receiving growing attention in the statistical literature.

Our findings are valuable in order to choose a profitable investment strategy basing financial decisions on transition probabilities and current regime classification. A constant update of the dynamic analysis through LMM may suggest the proper investment decision for the following month. Furthermore, it is possible to define a diversification strategy based on the different market indexes which are characterized by different risk-return profiles within the same latent state.

Our results represent the basis of an innovative classification of financial products achieved by exploiting the potential of statistical methodology for latent variables.

References

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, *41*, 164–171.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370.

De Angelis, L., & Paas, L. J. (2009). The dynamic analysis and prediction of stock markets through latent Markov model. *FEWEB Research Memorandum*, 2009-53. Amsterdam: Vrije Universiteit.

Dias, J. G., Vermunt, J. K., & Ramos, S. B. (2009). Mixture hidden Markov models in finance research. *Advances in data analysis, data handling and business intelligence*. Berlin: Springer.

Frühwirth-Schnatter, S., & Kaufmann, S. (2008). Model-based clustering of multiple time series. *Journal of Business and Economic Statistics*, *26*, 78–89.

Rossi, A., & Gallo, G. M. (2006). Volatility estimation via hidden Markov models. *Journal of Empirical Finance*, *13*, 203–230.

Hamilton, J. D., & Raj, B. (2002). *Advances in Markov-switching models*. Berlin: Springer.

Hamilton, J. D., & Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regimes. *Journal of Econometrics*, *64*, 307–333.

Otranto, E. (2008). Clustering heteroskedastic time series by model-based procedures. *Computational Statistics and Data Analysis*, *52*, 4685–4698.

Paas, L. J., Vermunt, J. K., & Bijmolt, T. H. A. (2007). Discrete time, discrete state latent Markov modelling for assessing and predicting household acquisitions of financial products. *Journal of the Royal Statistical Society A*, *170*, 955–974.

Rydén, T., Teräsvirta, T., & Åsbrink, S. (1998). Stylized facts of daily return series and the hidden Markov model. *Journal of Applied Econometrics*, *13*, 217–244.

Vermunt, J.K., & Magidson, J. (2007). *Technical guide for Latent GOLD 4.5: Basic and advanced*. Belmont, MA: Statistical Innovations Inc.

A MEM Analysis of African Financial Markets

Giorgia Giovannetti and Margherita Velucchi

Abstract In the last few years, international institutions stressed the role of African financial markets to diversify investors' risk. Focusing on the volatility of financial markets, this paper analyses the relationships between developed markets (US, UK and China) and some Sub-Saharian African (SSA) emerging markets (Kenya, Nigeria and South Africa) in the period 2004–2009 using a Multiplicative Error model (MEM). We model the dynamics of the volatility in one market including interactions from other markets, and we build a fully interdependent model. Results show that South Africa and China have a key role in all African markets, while the influence of the UK and the US is weaker. Developments in China turn out to be (fairly) independent of both UK and US markets. With the help of impulse-response functions, we show how recent turmoil hit African countries, increasing the fragility of their infant financial markets.

1 Introduction

With increasing globalization, world-wide integration of financial systems and crisis, international investors interest has been rekindled in African stock markets; it is well known, for example, the role of China in some emerging African markets.[1] Recently, "the Economist" characterized Africa as globalization's final frontier for investors (29/7/07), suggesting them to "Buy Africa" (19/2/2008). Indeed, before the global financial meltdown, African financial markets had experienced a large expansion in a very short time. The number of operating stock exchanges rose from just eight in 1989, to 23 in 2007, reaching a total market capitalization of over 2.1 billion US dollars. While small size and low liquidity remain a relevant aspect to be

[1] An extensive review of the institutional characteristics of the African stock markets appears in Irving (2005) and in Yartey (2008).

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_36,

further investigated, during the last few years, many African markets offered very large returns to investors.[2]

This paper analyzes the relationship between volatility in some developed financial markets (US, UK and China) and in emerging SSA markets (data availability restricted our analysis to Kenya, Nigeria and South Africa). We use daily data (composite indices for each market) for the period 2004–2009. Our aim is to test whether the financial turmoil that hit western markets spilled over into the African markets and how. We use a MEM (Multiplicative Error Model, Engle 2002) approach that allows us to describe and forecast the interactions and spill-over effects among different markets. We focus on the risk associated with financial markets, modelling the dynamics of the expected volatility of one market and including interactions with the past volatility from other markets. Then, we use a graphical representation of the spill-over effects and report impulse response functions (Engle et al. 2011). The impulse response functions allow us to describe how a shock in one market propagates, if at all, to other markets, and how long the effect is expected to last.

The paper is structured as follows: Section 2 outlines the econometric methodology, Sect. 3 presents the data and discusses results. Section 4 concludes.

2 Multiplicative Error Model for African Markets

We model the volatility of market i using its own past values and positive and negative news from other markets (Engle et al. 2011; Cipollini et al. 2009). From model estimation on the whole sample, we derive that SSA markets are interdependent and also depend on more financially developed markets to varying degrees. The relationships we find show how SSA African countries, contrary to some generalized investors' views, are not independent of global turmoil and help us to highlight the financial channels of transmission in SSA Africa. ME Models are a generalization of GARCH-type models estimated on non-negative valued processes (Engle 2002); the model we use in this paper is based upon Engle et al. (2011). Conditional on the information set I_{t-1}, the Multiplicative Error Model for market i is

$$sr_{i,t}|I_{t-1} = \mu_{i,t}\epsilon_{i,t},$$

where $\epsilon_{i,t}|I_{t-1} \sim Gamma(\phi_i, 1/\phi_i)$. Following Engle (2002), we know that a Gamma distribution assumption for ϵ is appropriate in this case. Given the unit expectation of the innovation term, $\mu_{i,t}$ is the conditional expectation of $sr_{i,t}$, where

[2] From the theoretical point of view, equity market integration plays a crucial role in development of infant financial markets. Finance theory suggests that an integrated stock market is more efficient than segmented national capital markets. Asset-pricing models also predict that integrated markets tend to respond more to global events than to local factors, although the reverse is also true (Errunza and Losq 1985). and Kim and Singal (2000) argue that a higher degree of market segmentation increases the level of risk, and this inevitably affects the local cost of capital, with ramifications for company financing and, in the long run, economic growth (Bracker et al., 1999).

$sr_{i,t}$ is a volatility proxy (range, squared returns, absolute returns, etc.). Its simplest specification is a base MEM(1, 1):

$$\mu_{i,t} = \omega_i + \beta_i \mu_{i,t-1} + \alpha_{i,i} sr_{i,t-1}. \quad (1)$$

This *base* specification can include other terms. In this paper we use the squared returns as a proxy of volatility ($sr_{i,t}$) and include (1) the lagged daily squared returns observed in other markets to link together different markets $sr_{j,t-1}$, $j \neq i$ (j represents developed markets: NYSE, FTSE and Shangai market), and (2) asymmetric effects in which the impact from own lagged volatility is split into two terms according to whether the lagged market returns ($r_{i,t-1}$) are negative and, respectively, positive (corresponding to dummy variables $D_{r_{i,t-1}<0}$, respectively, $D_{r_{i,t-1}>0}$). For each market we then use impulse response functions and dynamic forecasts according to Engle et al. (2011). The estimated models are used to analyse volatility shock propagation (Engle et al. 2011). For each market, we develop impulse response functions to describe how a shock in one market (an one-standard-deviation shock) may propagate to others. We report a graphical representation in which time (days since the shock hit the market) is on the horizontal axis and the volatility response (relative difference between a baseline and the response after the shock) is on the vertical axis. When a shock hits market i, the graph shows market volatility responses to the shock originating in that market; on average, the response of market i at time 0 to a shock in its own market will be higher than other markets' response, given the volatility persistence in the market. The graph shows the dynamic reactions (volatility spill-overs) of markets to single positive or negative shocks that may hit market i in a given day. Hence, the profiles turn out to be time dependent and may change as the originating market (the market which is shocked) and time (different shocks) vary.

3 Volatility Spillover in Selected SSA Financial Markets

For the markets at hand the descriptive statistics for daily absolute returns are reported in Table 1 showing the expected characteristics of financial data (leptocurtosis, asymmetry, volatility clustering).

Table 1 Descriptive statistics for the whole sample (2004–2009). Daily absolute returns. Obs: 1369

	China	Kenia	Nigeria	South Africa	UK	US
Mean	0.696	0.261	0.362	0.869	0.302	0.376
Median	0.188	0.031	0.085	0.219	0.046	0.046
Maximum	16.176	35.970	4.529	34.711	14.642	22.645
Minimum	0.000	0.000	0.000	0.000	0.000	0.000
Std. dev.	1.513	1.451	0.713	2.278	1.052	1.438
Skewness	5.335	17.279	3.286	8.125	8.443	9.136
Kurtosis	40.311	367.067	14.743	94.622	91.709	108.264

Figures 1 and 2 show the behavior of our six markets over the same period stressing the effect of boom and recent crisis.

Table 2 reports the estimates of the fully inter-dependent MEMs on three SSA financial markets (Kenya, Nigeria, South Africa) and China, UK and US for the whole period (1 January 2004–24 April 2009), separately for positive and negative shocks, which are likely to be spread differently. The results show that South Africa,

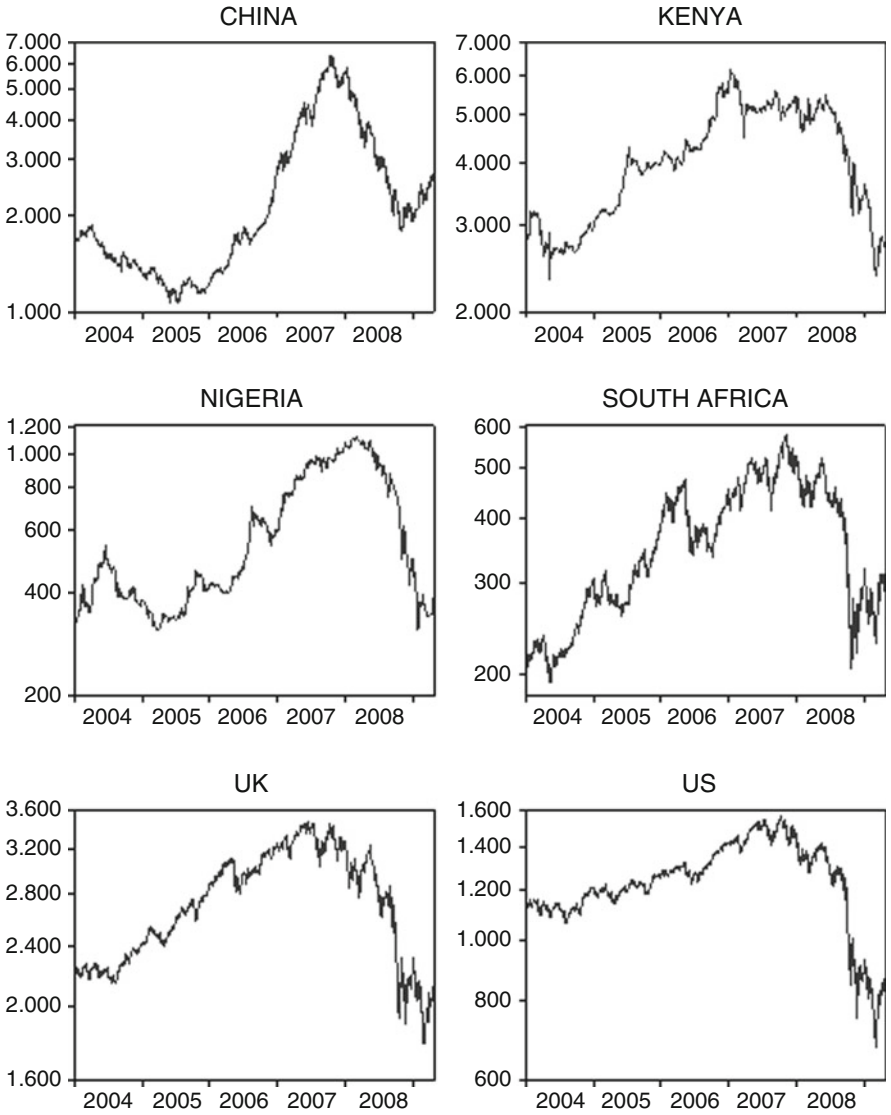


Fig. 1 Stock indices. January 2004–April 2009

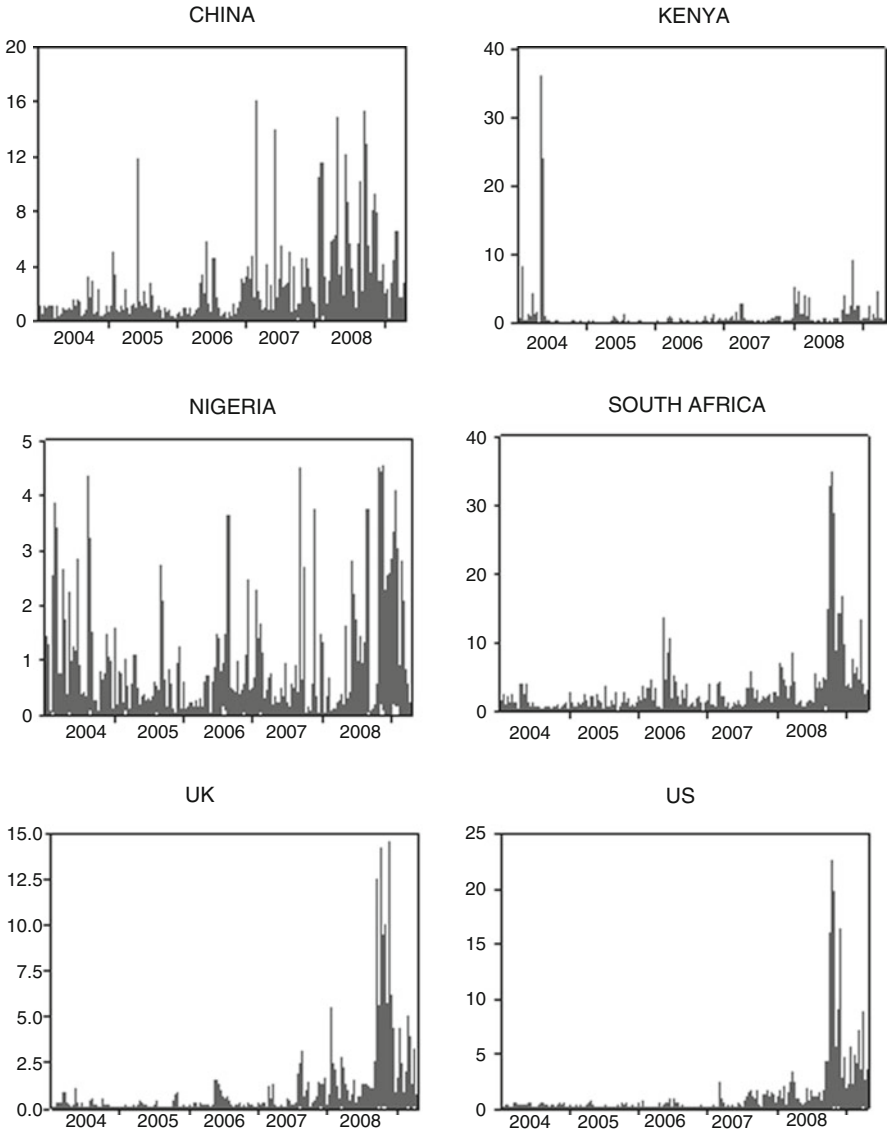


Fig. 2 Absolute returns. January 2004–April 2009. Figures are kept in different scale for ease of readability; markets, indeed, have very different volatility behaviors (see Nigeria, for instance)

China and the UK turn out to be the most influential markets (shocks, either positive or negative affect almost all markets). Nigeria and South Africa (but not Kenya) are hit by negative shocks in the US. Kenya, however, is influenced by positive shocks in the US. Negative shocks in South Africa affect the US but South Africa is only affected by the Nigerian market. There is no evidence of a significant relationship

Table 2 Estimated coefficients (t-stat in parenthesis). Sample: January 2004–April 2009. CORR(12) (respectively, CORRSQ(4)) is the LM test statistic for autocorrelation up to order 12 in the standardized residuals (respectively, squared standardized residuals)

	China	Kenya	Nigeria	South Africa	UK	US
ω	0.0134	0.0074	0.0811	0.0316	0.0008	0.0011
	[2.6143]	[3.5865]	[12.4384]	[3.3649]	[1.0196]	[1.7104]
μ_{t-1}	0.9255	0.6444	0.2256	0.8568	0.8919	0.9607
	[40.5247]	[13.3403]	[7.7779]	[24.5022]	[46.4033]	[88.0105]
ch^+	0.0494	−0.0015	0.0093	−0.0016	0.0058	0.0018
	[3.2503]	[−0.5910]	[1.3029]	[−0.1693]	[3.1574]	[1.2051]
ch^-	0.0716	−0.0052	0.0297	−0.0158	0.0012	0.0005
	[2.8248]	[−5.3302]	[3.4842]	[−2.6469]	[0.4349]	[0.3171]
ke^+	0.0096	0.2209	0.0001	0.0086	−0.0031	−0.0002
	[2.1164]	[2.6645]	[0.0102]	[0.5983]	[−0.8052]	[−0.1100]
ke^-	−0.0105	0.3032	0.0024	−0.0148	0.0000	−0.0024
	[−5.2380]	[4.4097]	[0.2574]	[−1.2033]	[0.0017]	[−1.3930]
ni^+	0.0094	−0.0037	0.5338	0.0254	0.0029	0.0019
	[1.0470]	[−1.1058]	[6.4341]	[1.5987]	[1.3018]	[0.7484]
ni^-	−0.0195	0.0019	0.4472	−0.0170	−0.0024	−0.0004
	[−3.0178]	[0.2942]	[6.9624]	[−1.2196]	[−1.4412]	[−0.1929]
sa^+	−0.0285	0.0447	−0.0047	−0.0171	0.0033	0.0022
	[−2.3113]	[2.3440]	[−1.1315]	[−0.8951]	[1.2055]	[0.7152]
sa^-	0.0040	0.0219	−0.0039	0.1038	0.0036	−0.0043
	[0.4008]	[2.4102]	[−2.2809]	[3.4972]	[1.3478]	[−2.4287]
uk^+	0.0224	0.0116	−0.0213	0.0642	−0.0262	−0.0104
	[0.3879]	[0.9081]	[−1.5579]	[0.7065]	[−1.2719]	[−0.4575]
uk^-	−0.0117	−0.0248	−0.0225	0.0829	0.0785	0.0389
	[−0.1803]	[−2.3505]	[−6.4670]	[0.8805]	[2.9896]	[2.4607]
us^+	−0.0162	−0.0355	0.0198	0.0187	−0.0115	−0.0412
	[−0.6294]	[−4.1995]	[1.3143]	[0.2564]	[−0.7519]	[−2.6338]
us^-	0.0571	0.0167	0.0413	0.1878	0.1022	0.0761
	[1.9125]	[1.4557]	[2.7151]	[2.6128]	[4.1099]	[3.4871]
loglik	−1577.8024	−582.6754	−1020.6835	−1595.6253	−581.9744	−663.2008
aic	2.3515	0.9145	1.5470	2.3773	0.9135	1.0308
bic	2.2986	0.8616	1.4941	2.3244	0.8606	0.9779
CORR(12)	3.7189	14.9780	17.7530	10.0810	10.2260	19.5490
	[0.9880]	[0.2430]	[0.2060]	[0.6090]	[0.5960]	[0.0760]
CORRSQ(4)	1.0015	0.5358	3.5965	15.2440	2.4136	1.8052
	[1.0000]	[1.0000]	[0.9900]	[0.2280]	[0.9980]	[1.0000]

between Nigeria and Kenya: the two markets seem to be fairly independent of each other. China has strong links with African markets, but it turns out to be independent of the UK and the US.

In Figs. 3–5, we report all markets volatility responses to a one standard deviation shock in market i. The representation uses the model estimates to derive a time-dependent profile that describes how volatility from one market (hit by a shock) may influence volatility behavior in other markets. To allow comparison among

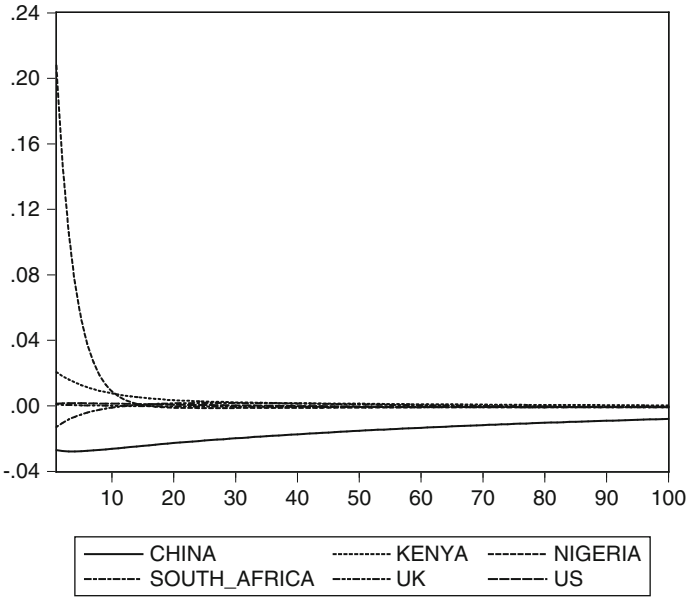


Fig. 3 Impulse response functions. We report time (days) in the horizontal axis and volatility on the vertical axis. Each *line* shows markets response to a one standard deviation shock originating in US (Jun., 29, 2006)

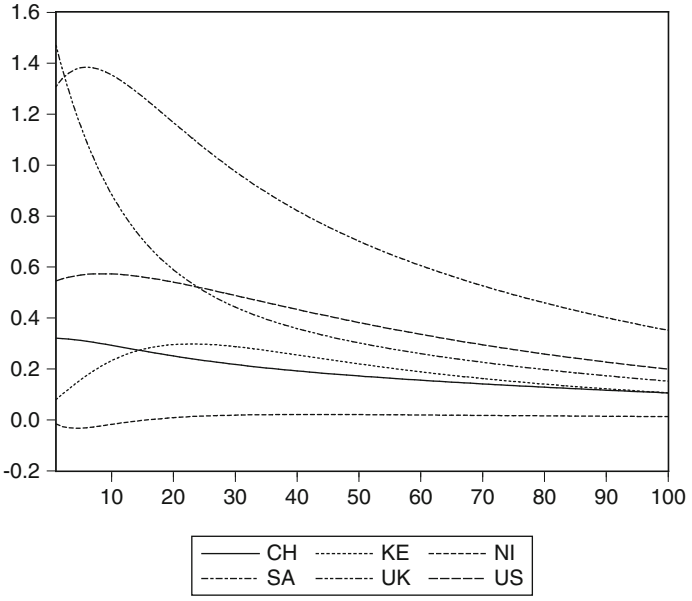


Fig. 4 Impulse response functions. We report time (days) in the horizontal axis and volatility on the vertical axis. Each *line* shows markets response to a one standard deviation shock originating in UK (Jan. 21, 2008)

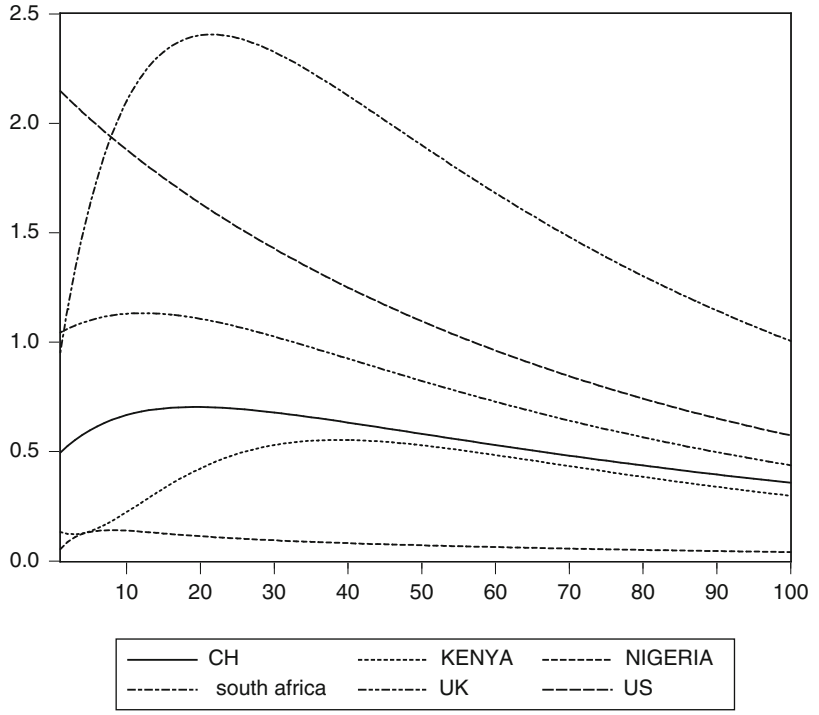


Fig. 5 Impulse response functions. We report time (days) in the horizontal axis and volatility on the vertical axis. Each *line* shows markets response to a one standard deviation shock originating in US (Sep., 15, 2008)

different market conditions, Fig. 3 reports the situation of a quiet day where all markets have low volatilities, and no negative or positive shocks hit any market. In contrast, Figs. 4 and 5 show the reaction in days of turmoil according to the estimation results. Figure 4 represents an example of impulse-response representation to a shock: on 21 January 2008, international stock markets suffered their biggest falls since 11 September 2001. The shock originated in the UK market, and then propagated to all markets; the US and South Africa over-react to the initial shock: in few days, their volatility response becomes higher (and larger) than the UK volatility response to its own shock. While China and Nigeria seem to be unaffected by the shock, Kenya shows a limited building effect, resulting from its increasing integration with western financial markets. In Fig. 5, we show how the collapse of Lehman Brothers on 15 September 2008 propagates from the US to all markets. It emerges clearly that this shock has a more relevant impact for all markets concerned: the scale of the impulse response is, indeed, almost double that of the previous example. The figure shows that, after about a week, South Africa volatility response is above that of the US; the UK also strongly reacts to the US shock; China and Kenya react more than in the previous example, while Nigeria seems to be insensitive even to

this large global shock. A further characteristic of the Lehman Brothers shock is that it is strongly persistent: after 3 months, it has not yet been re-absorbed. This is particularly evident in the case of South Africa, where the whole effect of the shock still persists.

4 Conclusive Remarks

Recent years have been characterized by major integration of international financial markets but African markets seem to have lagged behind this process, remaining fairly independent of international turmoil. Notwithstanding this, the recent global financial crisis has reached Africa, too, hitting some of the drivers of stock market development. In this paper, we focus on selected emerging SSA financial markets (Kenya, Nigeria, South Africa) volatility relationships with some developed markets (China, the UK and the US) using a Multiplicative Error approach to model and describe whether (and how) volatility spills over from one market to another. We focus on squared returns as a proxy of market volatility, and we model the dynamics of the expected volatility of one market including interactions with the past squared returns of the other markets. In doing so, we build a fully inter-dependent model that allows us to describe the relationships among the market volatilities. Impulse response functions are used to understand whether a shock originating in one market affects (and how) other markets. The results show that South Africa and China play a key role in all African markets, while the influence of events in the UK and the US is weaker and tend to be different for positive and negative shocks. Also, China turns out to be independent of the UK and the US. The impulse response representation shows that recent shocks in the UK and the US hit African financial markets hard. The SSA financial markets volatility response to a shock in either the UK or the US had cumulative effects that took time to be fully-developed and understood, thereby worsening their fragile economic conditions.

References

Bracker, K., Dockino, G., & Koch, P. (1999). Economic determinants of evolution in international stock market integration. *Journal of Empirical Finance*, *6*, 1–27.

Cipollini, F., Engle, R. F., Gallo, G. M., & Velucchi, M. (2009, April 3). *MEM based analysis of volatility in east Asia: Some lessons from financial crises*. Paper presented at Volatilities and Correlations in Stressed Markets Conference, Stern School of Business, New York University, New York.

Engle, R. F. (2002). New frontiers for ARCH models. *Journal of Applied Econometrics*, *17*, 425–446.

Engle, R.F., Gallo, G.M. & Velucchi, M. (2011), *A MEM-based analysis of volatility spillovers in east asian financial markets*, Review of Economics and Statistics, forthcoming.

Errunza, V., & Losq, E. (1985). International asset pricing under mild segmentation: Theory and test. *Journal of Finance*, *40*, 105–124.

Irving, J. (2005). *Regional integration of stock exchanges in eastern and southern Africa: Progress and prospects*. IMF Working paper WP/05/122.

Kim, J., & Singal, V. (2000). Stock market openings: Experience of emerging economies. *Journal of Business*, *73*, 25–66.

Yartey, C. A. (2008). *The determinants of stock market development in emerging economies: Is South Africa any different?* IMF Working Paper WP/08/32.

Group Structured Volatility

Pietro Coretto, Michele La Rocca, and Giuseppe Storti

Abstract In this work we investigate the presence of 'group' structures in financial markets. We show how this information can be used to simplify the volatility modelling of large portfolios. Our testing dataset is composed by all the stocks listed on the S&P500 index.

1 Introduction

The modelling of volatility for large dimensional portfolios is not an easy task and cannot be performed by standard methods. The need for modelling dynamic interdependencies among several assets would in principle require the estimation of an overwhelming number of parameters. Moreover, when the number of parameters is reasonable the implementation of standard inference procedures could still be not feasible even when the cross-sectional dimension is moderately large (say $\geq$50 assets). To simplify the modelling, it is usually necessary to introduce severe, often untested, homogeneity constraints on the dependence structure of the data. Recent literature has provided evidence in favour of the hypothesis that financial returns are characterized by cluster-wise dependence structures (see e.g., Bauwens and Rombouts 2007).

The aim of this paper is twofold. First, using data on all the S&P500 assets, model based cluster techniques will be applied in order to classify stocks into homogeneous risk categories. Second, we will investigate the economic implications of the clustering by mean of multivariate GARCH models. Namely, we will show that the dynamic properties of the conditional covariance matrix of stock returns are critically dependent on the volatility level. This provides a relevant argument in favour of the use of state-dependent multivariate conditional heteroskedastic models in order to predict portfolio volatility.

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_37,

2 Group-Structured volatility

The key assumption of this work is that the cross-sectional distribution of volatilities in a given market is a finite mixture of distributions. Each mixture component represents a group of assets that share similar risk behaviour. This hypothesis is based on the empirical assessment of the cross-sectional distribution of volatilities in the data, measured by the rolling standard deviation of asset returns, as described afterwards. If assets belongs to groups of homogeneous volatility it is natural to reduce the dimensionality of the portfolio aggregating stocks with similar volatility behavior in more than one sub-portfolio. In order to do that we need to cluster assets with respect to their volatilities and in order to do that we use model-based cluster analysis. A different approach to clustering financial time series has been proposed in (Otranto 2008) where the autoregressive metric (see, Corduas and Piccolo 2008) is applied to measure the distance between GARCH processes. The next section is devoted to a brief account of the cluster methods used in the present application.

2.1 Model-Based Clustering and Robustness

Model-based cluster analysis (see, Banfield and Raftery 1993) is a statistical tool used to investigate group-structures in the data. Finite mixtures of Gaussian distributions are a popular device used to model elliptically shaped clusters. In many situations it is sensible to approximate the distribution of the population under study with a mixture of Gaussian distributions, that is a convex combination of Gaussians. We interpret the mixture components as physically coinciding with the clusters in the population. For a wide class of mixture models estimation can be done by using maximum likelihood. Such estimates can be conveniently computed with algorithms of the EM class; and then, by simple probability arguments observed points can be assigned (classified) to the mixture components, that is clusters. In this paper we will focus mainly on Gaussian mixtures.

Mixture model estimation is seriously affected by small proportion of outlying observations. As we will see afterwards this issue is particularly relevant when modelling the cross-sectional distribution of asset volatilities. For a wide class of finite mixtures, including Gaussians, maximum likelihood estimates are not robust (see, Hennig 2004). This implies that a small proportion of outliers in the data could lead to poor estimates and clustering. One way to deal with this is to add a 'noise component', i.e., a mixture component that models the outliers. By noise we mean all those observations that are not consistent with the Gaussian cluster–prototype model assumed for the groups. There are several contributions in this direction (see, Banfield and Raftery 1993, Coretto 2008, Coretto and Hennig 2009). In his paper Hennig (2004) outlined a robust strategy in which the noise component is represented by a fixed improper density, which is a constant on the real line. He showed that the resulting estimates are robust to even extreme outliers, as opposed

to other noise component methods (in (Coretto 2008) these methods are compared extensively).

Coretto (2008) defined a pseudo-maximum likelihood estimator for such a model, called the robust improper maximum likelihood estimator (RIMLE), he studied its asymptotic behaviour and a computational strategy based on the EM algorithm. The RIMLE is based on an approximate representation of the population distribution by an improper density such as

$$\lambda_c(x;\eta) := \pi_0 c + \sum_{j=1}^{G} \pi_j \phi(x; m_j, v_j), \tag{1}$$

where $\phi(\bullet; m, v)$ is the Gaussian density with mean m and variance v, $\pi_0, \pi_1, \ldots, \pi_G$ are proportion parameters with $\pi_j > 0$ and $\sum_{j=0}^{G} \pi_j = 1$, finally $c \geq 0$ is a constant. The parameter vector η includes all proportions, means, and variances. For a fixed value of c the parameter vector of λ_c can be defined as the maximizer (under suitable constraints) of

$$l_n(\eta, c) := \sum_{i=1}^{n} \log \lambda_c(x_i, \eta), \tag{2}$$

where x_is are sample observations. The main issue about the RIMLE is how to get the value of c. This has been explored in Hennig (2004), Coretto (2008) and Coretto and Hennig (2009). The authors proposed statistical methods to fix the value of c based on optimality criteria.

After having obtained the RIMLE, $\hat{\eta}_n(c)$, we can assign observed points to mixture components by using Bayes' theorem. Let us introduce the following quantities

$$\tau_{ij} := \frac{\pi_0 c}{\lambda_c(x_i;\eta)} \text{ for } j = 0 \text{ and } \tau_{i,j} := \frac{\pi_j \phi(x_i; m_j, v_j)}{\lambda_c(x_i;\eta)} \text{ for } j = 1, 2, \ldots, G. \tag{3}$$

The quantity τ_{ij} can be interpreted as the posterior probability that the observation x_i belongs to the jth component. Since the density λ_c is not proper, we will refer to this as the improper posterior probabilities. In particular, the quantity τ_{i0} is the improper posterior probability that x_i is generated from the noise component. In model-based clustering we use these quantities in order to optimally define our clusters. That is, the observation x_i is assigned to the kth component if

$$k = \arg\max_{j=0,1,2,\ldots,G} \quad \hat{\tau}_{i,j},$$

where $\hat{\tau}_{i,j}$ is the estimate of τ_{ij} obtained by plugging-in $\hat{\eta}_n(c)$.

In Coretto (2008) and Coretto and Hennig (2009) the authors showed that when c is optimally fixed the RIMLE can achieve small misclassification rates independently of both the expected amount of noise (i.e., π_0) and the localization of the

noise support with respect to the support of the main clusters. The optimal choice of c provides very good results even in the presence of gross outliers.

2.2 Volatility Groups

Group-structures in asset volatilities could be related to heterogenous preferences that reflect in risk allocation. Moreover if the process that generates the returns includes regime switching from one level of volatility to another, one should observe that at a given date the entire market is composed of different groups of assets each sharing similar levels of volatility. The theoretical foundation of this assumption is still a work in progress.

We explored the cross-sectional distributions of the estimated volatility on the S&P500 based on daily returns. As a starting point for our research we used a rough volatility estimation based on rolling sample standard deviations. Let $r_{i,t}$ the log-return of the asset i at time t, we computed

$$\sigma_{i,t} = \sqrt{\frac{1}{L}\sum_{k=t-L}^{t-1}(r_{i,k} - \bar{r}_{i,t})^2}, \qquad \bar{r}_{i,t} = \frac{1}{L}\sum_{k=t-L}^{t-1} r_{i,k}.$$

We had good overall results for $L = 30$. In this work we assume that, at a given date t, the cross-sectional distribution of volatilities conditional on the information set at time $t-1$ is a finite mixture of Gaussian distributions, that is

$$\sigma_{i,t}|I_t \sim \mathrm{NM}(G, \theta_t), \qquad i = 1, 2, \ldots, S \tag{4}$$

where $NM(G, \theta_t)$ is the normal mixtures with G groups and parameter vector θ. Most of the cross-sectional distribution of volatilities presented small proportion of assets having extreme volatility. Since the analysis is done on a huge number of cross-sections identification of outlying observation would have been impossible to do if not based on an optimal 'automatic' statistical procedure such as the RIMLE. But the most important point about the use of the RIMLE in this work is that the RIMLE allow for smooth identification of noise/outliers. The latter means that we do not consider assets with abnormal volatility has extraneous to the population of interest, but we assign to a particular component of the population distribution (i.e., the noise component). If the noise component is not taken into account cross-sectional estimation of the model in (4) would deliver highly biased classifications. We approximated the model above by adding the improper noise component and for each date we estimated the corresponding RIMLE from the cross-section. The improper constant density is fixed using the 'filtering method' proposed and studied in (Coretto and Hennig 2009). Notice that the number of components is kept fixed across t. We fixed the number of groups to three plus the noise component. The latter will be discussed later as we move to the time series modelling. At each day

we obtained a classification of the portfolio's asset in four sub-portfolios. Hence in the classification step for each date we have three sub-portfolios characterized by different level of expected volatility and volatility dispersion, and a further group (usually very small) containing the outlying assets, that is assets with abnormal volatility.

3 Dynamic Properties of Clustered Volatilities

Aim of this section is to gain some insight on the economic interpretation of the results obtained through application of the above described model-based clustering procedure. To this purpose, we consider each cluster as a portfolio whose composition is possibly changing over time according to the volatility ranking of each asset. Within cluster (h), the weight associated to a given asset i at time t $(w_{i,t}^{(h)})$ is assumed to be proportional to the value of trades on that asset occurred at time $(t-1)$ $(v_{i,t-1}^{(h)})$:

$$w_{i,t}^{(h)} = \frac{v_{i,t-1}^{(h)}}{\sum_{j=1}^{n_{h,t}} v_{j,t-1}^{(h)}}$$

with $n_{h,t}$ being the total number of assets in cluster h at time t. The return on the portfolio associated to cluster h is then computed as $r_t^{(h)} = \sum_{j=1}^{n_{h,t}} w_{j,t}^{(h)} r_{j,t}^{(h)}$. The idea is to express the overall market return as a function of four indicators: returns on the low, medium and high volatility components of the market $(r_t^{(h)}$, $h = 1, 2, 3)$ and the noisy group $(r_t^{(4)})$. A Diagonal VECH model (Bollerslev et al. 1988) of order (1,1), abbreviated DVECH (1,1), is then fitted to the vector process which is generated by considering the returns on the first three components, which is $\underline{r}_t = (r_t^{(1)}, r_t^{(2)}, r_t^{(3)})'$.

$$\begin{aligned} \underline{r}_t &= \underline{\mu} + \underline{u}_t \\ \underline{u}_t &= H_t^{1/2} \underline{z}_t \qquad \underline{z}_t \underset{iid}{\sim} (\underline{0}, I_3) \end{aligned}$$

where $H_t = var(\underline{r}_t | I^{t-1})$. The DVECH(1,1) model implies that the elements of H_t evolve according to the following equation:

$$h_{ij,t} = \omega_{ij} + a_{ij} u_{t-1}^{(i)} u_{t-1}^{(j)} + b_{ij} h_{ij,t-1} \qquad i, j = 1, \ldots, h.$$

where $h_{ij,t} = cov(r_t, r_{j,t} | I^{t-1})$ and $u_t^{(i)} = r_t^{(i)} - \mu^{(i)}$. The reason for omitting the return on the noise component is twofold. First, we are interested in analyzing the structural components of the market. Second, for many time points the noise component either is not present at all, or there are on average no more than 2% of assets classified as noisy.

Table 1 ML estimates for the DVECH(1,1) parameters, with $\psi_{ij} = a_{ij} + b_{ij}$. All the estimates are significant at 0.01 level

	$\mu^{(i)}$	ω_{ij}	a_{ij}	b_{ij}	ψ_{ij}	$\sigma_{i,j}$	$\rho_{i,j}$
$h_{11,t}$	0.0947	0.0363	0.0627	0.8813	0.9439	0.6478	–
$h_{22,t}$	0.1389	0.1556	0.0962	0.8105	0.9067	1.6682	–
$h_{33,t}$	0.1894	0.6512	0.2418	0.6450	0.8868	5.7549	–
$h_{12,t}$	–	0.0595	0.0687	0.8622	0.9309	0.8611	0.8283
$h_{13,t}$	–	0.1226	0.0956	0.7884	0.8840	1.0571	0.5475
$h_{23,t}$	–	0.2472	0.1263	0.7484	0.8747	1.9721	0.6365

Table 2 P-values of the Wald tests of equality for the estimated DVECH (1,1) model's parameters

	a_{11}	a_{22}	a_{33}		a_{12}	a_{13}	a_{23}
a_{11}	–	0.0005	0	a_{12}	–	0.0063	0
a_{22}	–	–	0	a_{13}	–	–	0.0018
a_{33}	–	–	–	a_{23}	–	–	–
	b_{11}	b_{22}	b_{33}		b_{12}	b_{13}	b_{23}
b_{11}	–	0.0001	0	b_{12}	–	0.0005	0
b_{22}	–	–	0	b_{13}	–	–	*0.0444*
b_{33}	–	–	–	b_{23}	–	–	–
	ψ_{11}	ψ_{22}	ψ_{33}		ψ_{12}	ψ_{13}	ψ_{23}
ψ_{11}	–	0.0005	0.0005	ψ_{12}	–	0.0014	0
ψ_{22}	–	–	**0.2513**	ψ_{13}	–	–	**0.4727**
ψ_{33}	–	–	–	ψ_{23}	–	–	–

The estimates (Table 1) show that the dynamic properties of the estimated conditional covariances are clearly dependent on the volatility levels.

As expected, higher volatility levels imply higher expected returns. The value of the implied unconditional correlations ($\rho_{i,j}$) appears to be dependent on (a) the volatility gap between groups (b) the average (unconditional) volatility level of the involved groups.

Furthermore, in order to assess the significance of differences among the parameters of the marginal volatility models associated to each single group, we have performed a set of Wald tests whose p-values have been summarized in Table 2. The sensitivity to past shocks, as measured by a_{ii} ($i = 1, 2, 3$), appears to be directly related to the unconditional volatility level, tending to be higher in more volatile clusters. An opposite trend can be observed for the b_{ii} coefficients and for the persistence parameter ψ_{ii}. However the results of the Wald tests suggest that, while there is a significant persistence gap between groups 1 and 2, this is not true anymore as we focus on the comparison between groups 2 and 3. In practice, the negative variation in the inertia coefficient (b_{ii}) is compensating the opposite variation in the sensitivity to shocks parameter (a_{ii}). These results are consistent with previous findings (see e.g., Bauwens and Storti 2009 and references therein) suggesting that the

dynamic properties of the conditional variance of returns are dependent on the level of the volatility itself. In particular, it is well known that the persistence of volatility tends to be higher in tranquil rather than in turbulent market conditions.

A similar pattern characterizes the conditional covariance models. In line with what observed for the conditional variance, the sensitivity to shocks (a_{ij}) results to be positively related to the average unconditional volatility level of the involved groups. Namely, we observe a higher sensitivity parameter when we consider the conditional covariance between groups $\{1, 2\}$ than in all the other cases. Differently, as in the conditional variance case, the variation of the inertia coefficients (b_{ij}) appears to be characterized by an exactly opposite pattern. The results of the Wald tests (Table 2) confirm that all these differences are statistically significant. Also, we find that the conditional covariance between the less volatile groups $\{1, 2\}$ is characterized by a significantly higher persistence than what observed for the covariance between groups $\{2, 3\}$ and $\{1, 3\}$. No significant differences arise between the estimated persistences of $\{2, 3\}$ and $\{1, 3\}$.

As far as we know the literature on the presence of state-dependent features in the dynamics of conditional covariances is not as rich as the corresponding literature on conditional variances. Research has been mainly focused on the search for asymmetric dependences on the sign of past shocks (for a survey see, Storti 2008) but a systematic investigation of the relationship between conditional covariance dynamics and relevant state variables, such as volatility, is still missing.

References

Banfield, J., & Raftery, A. E. (1993). Model-based Gaussian and non-gaussian clustering. *Biometrics*, *49*, 803–821.

Bauwens, L., & Rombouts, J. (2007). Bayesian clustering of many GARCH models. *Econometric Reviews*, *26*, 365–386.

Bauwens, L., & Storti, G. (2009). A component GARCH model with time varying weights. *Studies in Nonlinear Dynamics and Econometrics*, *13*, online at http://www.bepress.com/snde/vol13/iss2/art1.

Bollerslev, T., Engle, R. F., & Wooldridge, J. M. (1988). A capital asset pricing model with time-varying covariances. *Journal of Political Economy*, *96*(1), 116–131.

Corduas, M., & Piccolo, D. (2008). Time series clustering and classification by the autoregressive metric. *Computational Statistics and Data Analysis*, *52*(4), 1860–1872.

Coretto, P. (2008). The noise component in model-based clustering. Ph.D. thesis, University College London (Advisor: Dr. C. Hennig), London, UK.

Coretto, P., & Hennig, C. (2009). Maximum likelihood estimation of heterogeneous mixtures of gaussian an uniform distributions. Article submitted, available upon request.

Coretto, P., & Hennig, C. (2009). Robust model-based clustering using an improper constant density selected by a distance optimization for non-outliers. In *JSM proceedings, IMS general theory section* (pp. 3847–3858). Alexandria, VA: American Statistical Association.

Hennig, C. (2004). Breakdown points for maximum likelihood estimators of location scale mixtures. *The Annals of Statistics*, *32*(4), 1313–1340.

Otranto, E. (2008). Clustering heteroskedastic time series by model-based procedures. *Computational Statistics and Data Analysis*, *52*(10), 4685–4698.

Storti, G. (2008). Modelling asymmetries in conditional correlations by multivariate BL-GARCH models. *Statistical Methods and Applications*, *17*, 251–274.

Part VIII
Functional Data Analysis

Clustering Spatial Functional Data: A Method Based on a Nonparametric Variogram Estimation

Elvira Romano, Rosanna Verde, and Valentina Cozza

Abstract In this paper we propose an extended version of a model-based strategy for clustering spatial functional data. The strategy, we refer, aims simultaneously to classify spatially dependent curves and to obtain a spatial functional model prototype for each cluster. The fit of these models implies to estimate a variogram function, the trace variogram function. Our proposal is to introduce an alternative estimator for the trace-variogram function: a kernel variogram estimator. This works better to adapt spatial varying features of the functional data pattern. Experimental comparisons show this approach has some advantages over the previous one.

1 Introduction

In the last years many approaches for clustering functional data have been proposed (Abraham et al. 2005; Heckman and Zamar 2000; James and Sugar 2005; Romano 2006). Most of them are based on the assumption of independence between curves. However, since in many applicative fields this assumption does not hold, there is a real need to introduce new statistical methods to deal with this problem. A very recent contribution in this framework is a model based strategy (Romano et al. 2010).

The method is a special case of Dynamic Clustering (Diday 1971). It aims to obtain a partition of the curves and to identify a spatial functional model for each cluster by locally minimizing the spatial variability among the curves. The prototype, in this sense, is a curve spatially predicted, obtained as a pointwise linear combination of the smoothed data. The prediction problem is solved by estimating a linear concurrent model (Ramsay and Silverman 2005) in a spatial framework for each cluster, where, the predictor as the same expression of an ordinary kriging predictor in unsampled location of the space. The computation of these kriging functions implies an estimate of a variogram function: the trace variogram function (see, for instance, Delicado et al. 2007). This estimation is usually based on the method-of-moments, however, there are many contexts for which it does not collect all spatial information present in the pattern. In this paper we introduce an estimator

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_38,

for the trace-variogram function. Especially we adapt the kernel based method for non paramentric variogram estimation proposed by Keming et al. (2007) in the functional setting. It is a kernel estimator, a more general weighted average than the classical estimator and it is robust in the sense that nearest-neighbor parameter selection is distribution free.

The paper is organized as follows. Sect. 2 presents a kernel estimator for the variogram function when data are curves. Sect. 3 describes the model based clustering strategy with non parametric trace-semivariogram estimation. Quality and performance of the method with and without the introduction of the new variogram estimator are presented in Sect. 4.

2 A Kernel Based Method for Nonparametric Variogram Estimation when Data are Curves

Let $\{\boldsymbol{\chi}_{\mathbf{s}}(t) : t \in T, s \in D \subset R^d\}$ be a random field, the set $D \subset R^d$ is a fixed subset of R^d with positive volume and $\boldsymbol{\chi}_{\mathbf{s}}$ 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. Moreover we assume to 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. Thus,

- $E(\boldsymbol{\chi}_{\mathbf{s}}(t)) = m(t)$, for all $t \in T,\ s \in D$.
- $V(\boldsymbol{\chi}_{\mathbf{s}}(t)) = \sigma^2$, for all $t \in T,\ s \in D$.
- $Cov(\chi_{s_i}(t), \chi_{s_j}(t)) = C(h,t)$ where $h_{ij} = \left\| s_i - s_j \right\|$ 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 $h_{ij} = \left\| s_i - s_j \right\|$ and all $s_i, s_j \in D$.

The function $\gamma(h,t)$ as function of h is called semivariogram of $\chi(t)$.

A general problem in Spatial statistics is to estimate the spatial correlation or equivalently the variogram of an intrinsic stationary process. Most natural empirical estimators of the variogram cannot be used for this purpose, as they do not achieve the conditional negative-definite property. Thus several variogram estimator are proposed.

In the functional framework a first attempt was carried out (Delicado et al. 2007). It is an extension of the classical estimator CL (Cressie 1993) obtained by the method of the moment by replacing the summation by an integral.

The classical variogram estimator is given by:

$$\hat{\gamma}(h) = \frac{1}{|2N(h)|} \sum_{i,j \in N(h)} \int_T \left(\chi_{s_i}(t) - \chi_{s_j}(t)\right)^2 dt \tag{1}$$

where $N(h) = \{(s_i; s_j) : \|s_i - s_j\| = h\}$ and $|N(h)|$ is the number of distinct elements in $N(h)$.

As with the sample variance, if the data are contaminated either by outliers or severe skewness, then the classical variogram estimator may be affected. In other words, it is not robust to departures from normality.

In geostatistics this has led to some extensions of the classical variogram estimator. Some of these focus on the problem of replacing the distance h with a more general distance.

Our goal is to generalize one of these estimators in the spatial functional framework. In particular we extend the variable nearest-neighbor estimator VNN proposed by Keming et al. (2007) for process with stationary of the second order.

The VNN estimator is based on a non-constant nearest-neighbor parameter which is able to adapt to many situations such as for example continuity over the range of data distance, non decreasing behavior over function with the increasing of distance h. This estimator is expressed by:

$$\hat{\gamma}_{VNN}(h) = \frac{\sum_{i<j} \frac{1}{\delta_0(s_{ij})} K\left(\frac{h - s_{ij}}{\delta\delta_0(s_{ij})}\right) Z_{ij}}{\sum_{i<j} \frac{1}{\delta_0(s_{ij})} K\left(\frac{h - s_{ij}}{\delta\delta_0(s_{ij})}\right)} \tag{2}$$

where the function $\delta_0(s_{ij}) > 0$ depends on the locations through the distance $s_{ij} = |s_i - s_j|$, $K(.)$ is a kernel function generally symmetric respect to the mean, including the indicator function $I(\cdot)$ which is defined on the radius $\delta\delta_0$ and $Z_{ij} = (Z_i - Z_j)^2$, where Z is a partial realization of a spatial process. If the process is known to be stationary of the second order, the VNN estimator, obtained by setting $\delta_0(s_{ij}) = 1$, is expressed by:

$$\hat{\gamma}_{VNN}(h) = \sum_{i<j} w_{ij} (Z_i - Z_j)^2 \tag{3}$$

where w_{ij} are the weights are obtained by setting $K_{\delta\delta_0(s_{ij})}(h) = \frac{1}{\delta_0(s_{ij})} K\left(\frac{h}{\delta\delta_0(s_{ij})}\right)$, so that:

$$w_{ij} = \frac{K_{\delta\delta_0(s_{ij})}(h - \|s_i - s_j\|)}{\sum_{i<j} K_{\delta\delta_0(s_{ij})}(h - \|s_i - s_j\|)} \tag{4}$$

Since we assumed that the mean function is constant

$$V(\chi_{s_i}(t) - \chi_{s_j}(t)) = E\left[(\chi_{s_i}(t) - \chi_{s_j}(t))^2\right] \tag{5}$$

so using the Fubini's theorem $\hat{\gamma}(h) = \frac{1}{2} E\left[\int_T (\chi_{s_i}(t) - \chi_{s_j}(t))^2 dt\right]$

Then the *VNN* estimator for the quantity above is given by

$$\hat{\gamma}_{VNN}(h) = \sum_{i<j} w_{ij} \int_T (\chi_{s_i}(t) - \chi_{s_j}(t))^2 \, dt \tag{6}$$

that is

$$\hat{\gamma}_{VNN}(h) = \frac{\sum_{i<j} \frac{1}{\delta_0(s_{ij})} K\left(\frac{h-s_{ij}}{\delta\delta_0(s_{ij})}\right) \int_T \left(\chi_{s_i}(t) - \chi_{s_j}(t)\right)^2 dt}{\sum_{i<j} \frac{1}{\delta_0(s_{ij})} K\left(\frac{h-s_{ij}}{\delta\delta_0(s_{ij})}\right)} \tag{7}$$

Note that the propriety of the original estimator is the same of as the one proposed, with some variation about the convergence since we are dealing with functional data.

Such as in the geostatistics framework, the introduction of the variable nearest-neighbor parameter $\delta_0 s_{ij}$ gives a flexibility and a higher spatial adaptation in estimating the variogram. The estimation here expressed is nonparametric and robust in the sense that the selection of the $\delta\delta_0$ selection is distribution free.

In the following we explain the model based clustering strategy and the use of the kernel variogram estimator.

3 Dynamic Clustering for Spatial Functional Data: A Model Based Approach

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 $P = \{P_1, \ldots, P_C\} \in \mathscr{P}_c$ (where $\mathscr{P}_c$ is the family of all the partition $\{P_1, \ldots, P_c\} \in \mathscr{P}_C$ in C clusters) and a set $G = \{g_1, \ldots, g_C\} \in \mathscr{G}_c$ (where $\mathscr{G}_c$ is the family of all admissible representation of C clusters prototypes) such that the criterion of best fitting between G and P is minimized

$$\Delta(P^*, G^*) = Min\{\Delta(P, G) \mid P \in \mathscr{P}_c, \in \mathscr{G}_c\} \tag{8}$$

Since we deal with spatial functional data, the criterium to optimize is:

$$\Delta(P, G) = \sum_{c=1}^{C} \sum_{i \in P_c} \int_T V\left(\chi_{s_i}(t) - \chi_{s_c}(t)\right) dt \tag{9}$$

where $V\left(\chi_{s_i}(t) - \chi_{s_c}(t)\right)$ is the spatial variability.

We assumed that data are generated from a functional linear concurrent model (Ramsay and Silverman 2005) so the criterion can be written

$$\Delta(P, G) = \sum_{c=1}^{C} \sum_{i \in P_c} \int_T V\left(\chi_{s_i}(t) - \sum_{i=1}^{n_c} \lambda_i \chi_{s_i}(t)\right) dt \qquad u.c. \sum_{i=1}^{n_c} \lambda_i = 1 \tag{10}$$

where n_c is the number of the elements in each cluster and the prototype $\chi_{s_c} = \sum_{i=1}^{n_c} \lambda_i \chi_{s_c}(t)$ is an ordinary kriging predictor for curves in the clusters. According to this criterion the kriging coefficients represent the contribution of each curve to the prototype estimate in an optimal location s_c.

Thus, the parameters to be estimated are: the kriging coefficients, the spatial location of the prototypes, the residuals spatial variance for each cluster.

For a fixed value of the spatial location of the prototype s_c, this is a constrained minimization problem, due to the unbiasedness constraint.

In order to obtain a solution of the optimized problem, it is necessary to solve a linear system by means of Langrange multiplier method. In matrix notation, by considering the relation $\gamma_{rs}(t) = \sigma^2(t) - C_{rs}(t)$, it can be expressed as:

$$\begin{pmatrix} \int_T \gamma_{s_1 s_1}(t)dt & \dots & \int_T \gamma_{s_1 s_{n_c}}(t)dt & 1 \\ \vdots & \ddots & \dots & \dots \\ \int_T \gamma_{s_{n_c} s_1}(t)dt & \dots & \int_T \gamma_{s_1 s_{n_c}}(t)dt & 1 \\ 1 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ -\mu \end{pmatrix} = \begin{pmatrix} \int_T \gamma_{s_0 s_1}(t)dt \\ \vdots \\ \int_T \gamma_{s_0 s_n}(t)dt \\ 1 \end{pmatrix}$$

Where the function $\gamma(h) = \int_T \gamma_{s_i s_j}(t)dt$, $h = \left\| s_i - s_j \right\|$ is the so called trace-semivariogram function. In order to solve the system above, an estimator of the trace-semivariogram is needed. In the original clustering approach we use the classical estimator obtained thorough an adaptation of the classical method of the moment. Here we propose to use the kernel-based estimator proposed above. 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. Therefore 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:

$$\int_T V\left(\chi_{s_i}(t) - \sum_{i=1}^{n_c} \lambda_i \chi_{s_i}(t)\right) dt = \sum_{i=1}^{n_c} \lambda_i \int_T \gamma_{s_i, s_c}(t)dt - \mu \quad u.c. \sum_{i=1}^{n_c} \lambda_i = 1 \tag{11}$$

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.

We use this measure to compare the clustering results obtained with the introduction of the new estimator.

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 to observe the spatial functional data on a grid where unsampled locations are observed. These are candidates for the location of the prototype. In this sense the location is a parameter

that must be estimated and the objective function may have several local minima correspondent to different local kriging. We propose to solve this problem evaluating, for each cluster, local kriging on unsampled locations. The prototype is the best predictor in terms of the best spatial functional fitting (5) among the set of estimates on the unsampled locations of the grid.

Moreover the spatial coordinates s_c of the prototype χ_{s_c} are chosen among a set S^* of possible locations in order to minimize the spatial variance.

Once we have estimated the prototypes we allocate each new curve to the cluster according to the following *allocation* function: $\kappa = \chi_s \longmapsto \mathscr{P}_c$ It allows to assign χ_s to cluster c of P_c $\kappa(G) = P = \{P_1, \ldots, P_C\}$, according to the minimum-spatial variability rule:

$$P_c := \{i \in \chi_s : \delta(\{i\}, \chi_{s_c}) \leq \delta(\{i\}, \chi_{s_{c*}}) \; for \; 1 \leq c^* \leq C\} \tag{12}$$

with $\delta(\{i\}, \chi_{s_c}) = \frac{1}{\lambda_\alpha} \int_T V\left(\chi_{s_i}(t) - \chi_{s_c}(t)\right) dt$ where λ_α is the kriging coefficient or weight such that $|s_\alpha - s_c| \cong h$ where $h = |s_i - s_c|$. These weights are derived based on a data-driven weighting function making the sum of the weights equal to one, thereby reducing the effect of bias towards input sample values. Note,however, that it is possible that some may be negative. A negative block can be assigned a zero value to avoid problems in post kriging analyses, such as pit-optimization (Stein 1999). Applying iteratively the assignment function followed by the allocation function under some conditions 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.

4 Dealing with Real Data

In this study, our aim is to assess the effectiveness of the kriging predictor by introducing the kernel trace-semivariogram estimation and its accuracy in the clustering process for the estimation of the spatial functional prototype.

This is performed by comparing: the goodness of kriging prediction of the VNN estimator with reference to the classical trace-variogram estimation; the clustering results obtained by introducing the *VNN* and *CL* estimator.

A real dataset which stores the curves of sea temperature along several locations of the Italian Coast provides the object of our analysis(see http://www.mareografico.it). The mareographic Network is composed by 26 survey stations distributed across the Italian territory. For each location, we have the recording of two weeks of data. Moreover the spatial coordinates (latitude, longitude) are available.

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.

As first stage, we use a cross-validation method to verify the goodness of fit of kriging prediction model with the two different trace-semivariogram estimators and then we compare the obtained results. Especially after that each curve has been

removed and further predicted from remaining data, we observe the distribution of the global uncertainty measure for kriging prediction (GUMK) on the estimated locations. Note that a spherical model was fitted to estimate the trace-variogram. Table 1 shows the mean and the standard deviation of the *GUMK*. It indicates that the predicted curves with the kernel variogram estimator has a less variance and consequently better goodness. We then evaluate the clustering results by performing two different *DCA*. The first one is based on the classical trace-semivariogram estimator, the second on the *VNN* estimator. Our task is to find the suitable prototype and their corresponding clusters. To get an initial partition of data into spatially contiguous regions, we run, for both strategies, a standard k-means algorithm on the spatial locations of the observed data. Since the proposed method, detects the prototypes of each cluster starting from a regular spatial grid, we set the number of cells the grid is made of. In our experiments, several sizes of grid have been tested. We compare the value of the optimized criterion obtained for the two methodologies after running the algorithms with a variable number of clusters $C = 2, 3, 4$. Reading Table 2 it is interesting to note that the DCA_{VNN} gives better fitting than DCA_{CL} for all the number of cluster evaluated. From a technical point of view, looking at the clustering structure in both the clustering results the obtained partition divides the Italian coast into three homogeneous zones representing three macro areas of the sea, respectively: Tirreno sea, Adriatico sea and Ligure sea. The clusters obtained by DCA_{CL} contain respectively $12, 9, 5$ elements with prototypes located in: Sorrento, Francavilla Al Mare and Alassio (Table 3). While clusters obtained by DCA_{VNN} contain $10, 10, 6$ elements and prototypes located in: Napoli, Pescara, Pietra Ligure

Table 1 Mean and Std value of GUMK

Krig.Pre	Mean	StD
By VNN	0.013	0.0007
By Trad.	0.122	0.105

Table 2 Criterion

Δ	$C = 2$	$C = 3$	$C = 4$
DCA_{VNN}	1990	1328	1128
DCA_{CL}	2021	1400	1200

Table 3 Locations of the prototypes by *DCA*

Prototypes by DCA_{CL}	Latitude	Longitude
Sorrento	$40°37'53.98''$	$14°21'52.69''$
Francavilla Al Mare	$42°25'14.26''$	$14°17'15.91''$
Alassio	$44°00'10.58''$	$8°09'34.32''$
Prototypes by DCA_{VNN}	Latitude	Longitude
Napoli	$40°50'60''$	$14°15'80''$
Pescara	$42°27'57''$	$14°12'52''$
Pietra Ligure	$44°08'57''$	$8°17'29''$

(Table 3). These results provide evidence that some curves have migrated from one cluster to an other. This reflects the effect of considering the VNN estimator in the prototype computation. This is clearly highlighted by observing that weight of each curve on the prototypes computation has changed and the prototypes are located in different spatial sites Table 1. Especially, according to the DCA_{CL} results, in the first cluster the greatest weight $\lambda_1 = 0.63$ corresponds to Salerno; in the second cluster, the greatest weight $\lambda_3 = 0.49$ corresponds to Ortona; in the third cluster, the greatest weights $\lambda_2 = 0.3, \lambda_5 = 0,29$ correspond to Genova and La Spezia. While according to the DCA_{VNN} results, in the first cluster the greatest weight $\lambda_4 = 0.55$ corresponds to Sorrento; in the second cluster, the greatest weight $\lambda_2 = 0.53$ corresponds to Vieste; in the third cluster, the greatest weight $\lambda_3 = 0.4$ corresponds La Spezia.

References

Abraham, C., Corillon, P., Matnzer-Löber, E., & Molinari, N., (2005). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics*, *30*, 581–595.

Cressie, N.A.C. (1993). *Statistics for spatial data*. New York: Wiley.

Diday, E. (1971). La Méthode des nuées dynamiques. *Review the Statistics Applications*, *XXX*(2), 19–34.

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.

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.

Keming, Y., Mateu, J., & Porcu, E. (2007). A kernel-based method for nonparametric estimation of variograms. *Statistica Neerlandica, Netherlands Society for Statistics and Operations Research*, *61*(2), 173–197.

Ramsay, J.E., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). New York: Springer.

Romano, E. (2006). *Dynamical curves clustering with free knots spline estimation*. PhD Thesis. Naples: University of Federico II.

Romano, E., Balzanella, A., & Verde, R. (2010). Clustering spatio-functional data: A model based approach. In *Proceedings of the 11th IFCS biennial conference and 33rd annual conference of the Gesellschaft für Klassifikation e.V., Dresden, March 13–18, 2009 Studies in classification, data analysis, and knowledge organization*. Berlin-Heidelberg, New York: Springer. ISBN: 978-3-642-10744-3.

Stein, M.L. (1999). *Interpolation of spatial data: Some theory for kriging*. New York: Springer-Verlag.

Prediction of an Industrial Kneading Process via the Adjustment Curve

Giuseppina D. Costanzo, Francesco Dell'Accio, and Giulio Trombetta

Abstract This work addresses the problem of predicting a binary response associated to a stochastic process. When observed data are of functional type a new method based on the definition of special Random Multiplicative Cascades is introduced to simulate the stochastic process. The *adjustment curve* is a decreasing function which gives the probability that a realization of the process is *adjustable* at each time before the end of the process. For real industrial processes, this curve can be used for monitoring and predicting the quality of the outcome before completion. Results of an application to data from an industrial kneading process are presented.

1 Introduction

In this paper we deal with the problem of predicting a binary response associated to a stochastic process, namely an industrial one. This problem arises when the industrial processes can be described in terms of a continuous phenomenon evolving in a certain interval of time $[0, T]$ and resulting in an outcome not observable before the completion of the process itself. Such an outcome, in accordance with some target values related to the process, can, in simple terms, expressed as negative or positive, bad or good and so on. For various reasons (for example economical ones) it could be useful to try to anticipate the outcome before the completion of the process: if we were able to predict the realization at the time T of the outcome (i.e., bad or good) earlier (at $t < T$) than the end of the observed process, this would enable us to direct the process in order to achieve the optimal target value or stop any process resulting in an undesired bad outcome. The situation we have in general illustrated is common in many real industrial applications requiring process control (see Box and Kramer 1992 for a discussion; for several examples see amongst Kesavan et al. 2000). In our case, the situation depicted in Fig. 1 is considered where for each type of flour, during the kneading process the resistance of dough (density) in a interval of time $[0, T]$ has been recorded. The achieved dough resistance in T affects the outcome of the process, that is the quality – good or bad – of the resulting cookies. The obtained curves can be then used to predict the quality of cookies made

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_39,

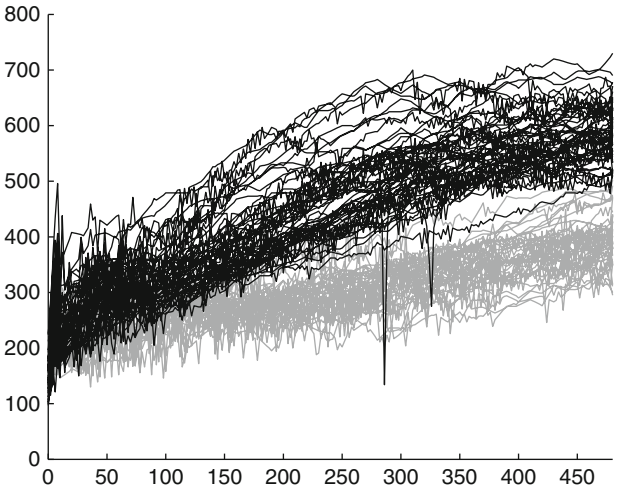


Fig. 1 Danone's original data

with this dough. In fact, if the cookies quality could be predicted in advance, remedial action could be taken to improve or stop the production of bad quality cookies. When predictors are of functional type and response is a categorical variable Y defining K groups, $K \geq 2$ various linear regression and classification methods have been proposed in literature (e.g., see Escabias et al. 2007, James 2002, Preda and Saporta 2005, Saporta et al. 2007). To address the problem of *anticipated prediction* of the outcome at the time T of the process in $[0, T)$ in Costanzo et al. (2006), we measured the predictive capacity of a linear discriminant analysis (LDA) for functional data model on the whole interval $[0, T]$. Then, depending on the quality of prediction,we determined a time $t^* < T$ such that LDA model considered on $[0, t^*]$ gives similar results, in terms of prediction of the outcome, as that considered on $[0, T]$. In this work we consider instead a new tool: the *adjustment curve* for a random binary response associated to a stochastic process, that is a decreasing function which gives the probability that a new realization of the process is *adjustable* at each time $t \in [0, T]$. The paper is organized as follows. In Sect. 2 we present the notion of adjustment curve. In Sect. 3 we briefly describe our method based on the definition of special Random Multiplicative Cascades (RMC) used to derive the adjustment curve of the stochastic process. Sect. 4 is devoted to the results of our case study.

2 Adjustment Curve for the Binary Response of a Real Process

We start by considering a matrix of data of functional type (see Ramsay et al. 2005) $FD = \left[x^i\left(t_j\right)\right], i = 1, \ldots L,\ t_j = j \cdot \frac{T}{S}$ for $j = 0, \ldots, S$, where each row represents the values of a continuous curve observed at discrete times $t_j, j = 0, \ldots, S$.

Next, we consider, the column vector $R = (r^i),\ i = 1, \ldots, L$, where r^i is a binary outcome associated to the row $\left(x^i\,(0) \ldots x^i\left(t_j\right) \ldots x^i\,(T)\right)$ for $i = 1, \ldots, L$; for example $r^i \in \{bad, good\}$. Let us suppose that FD and R jointly arise from a continuous phenomenon which can be simulated by a couple (X, Y), where we assume X is a stochastic process whose realizations are real continuous functions with $\{x(0) : x \in X\} = \left\{x_0^i : i = 1, \ldots, L\right\}$, linear on the intervals $[t_j, t_{j+1}]$, $t_j = j \cdot \frac{T}{S}$ for $j = 0, \ldots, S-1$, with a constraint on the increment, i.e., $|x(t_{j+1}) - x(t_j))| \leq M(x, j)$ for $j = 0, \ldots, S-1$. In the simplest case we can assume that the increment does not exceed a certain mean constant value obtainable from the real data, i.e., $M(x, j) = M$ for each $x \in X, j = 0, \ldots, S-1$. We will denote by $\mathscr{S}$ the class of such stochastic processes, $X = \{X\,(t)\}_{t\in[0,T]}$ a stochastic process in $\mathscr{S}$, $x = x\,(t)$ $(t \in [0, T])$ a realization of X. Moreover $\mathscr{R}$ will denote the class of all binary responses Y associated to X. Without loss of generality we can assume $Y \in \{bad, good\}$. In order to introduce the definition of the adjustment curve $\gamma_{a,D} : [0, T] \to [0, 1]$ for the binary outcome R of the functional data FD, we require that the condition $x^i\,(T) < x^j\,(T)$ for each $i : r^i = bad$ and $j : r^j = good, i, j = 1, \ldots, L$ is satisfied. That is, we assume there exist a value $X(T) \in \mathbb{R}$ such that $r^i = bad$ if, and only if, $x^i\,(T) < X(T)$ and $r^i = good$ if, and only if, $x^i\,(T) \geq X(T)$, $i = 1, \ldots, L$. For each $i = 1, \ldots, L$, let $s_D^i : [0, T] \to \mathbb{R}$ be the piecewise linear function whose node-set is $N^i = \left\{(t_j, x^i\,(t_j)) : t_j = j \cdot \frac{T}{S}, j = 0, \ldots, S\right\}$. Define:

$$b_D(j) = \max\left\{x^i(t_j) : i = 1, \ldots, L \text{ and } r^i = bad\right\} \quad (j = 0, \ldots, S) \tag{1}$$

$$g_D(j) = \min\left\{x^i(t_j) : i = 1, \ldots, L \text{ and } r^i = good\right\} \quad (j = 0, \ldots, S). \tag{2}$$

Let $i \in \{1, \ldots, L\}$ and $r^i = bad$ $(r^i = good)$: the piecewise linear interpolant s_D^i is called adjustable at the time $t \in [0, T]$ (for short $t-adjustable$) if there exists $t_j \geq t$ with $j \in \{0, 1, \ldots, S\}$ such that $s_D^i(t_j) \geq g_D(j)$ $(s_D^i(t_j) \leq b_D(j))$. The adjustment curve $\gamma_{a,D} : [0, T] \to \mathbb{R}$ for the binary outcome R of the functional data FD is the function

$$\gamma_{a,D}(t) = \frac{\left|\left\{s_D^i : i = 1, \ldots, L \text{ and } s_D^i \text{ is } t - adjustable\right\}\right|}{L} \quad (t \in [0, T]).$$

Given a set of curves deriving from a real continuous process the *adjustment curve* is a decreasing step function which gives the relative frequency of curves adjustable (with respect to the final outcome in T) at each time $t \in [0, T]$. As a consequence, the complementary curve $1 - \gamma_{a,D}(t)$ gives, at each time, the relative frequency of the curves that are *definitively* good or bad. Further, we observe that by the two data sets (1) and (2) it is possible to deduce the binary response associated to $s_D^i, i \in \{1, \ldots, L\}$ at each time $t_j, j = 0, \ldots, S-1$, before time T; indeed: if $s_D^i(t_j) > b_D(j)$ then $r^i = good$ or if $s_D^i(t_j) < g_D(j)$ then $r^i = bad$; otherwise r^i is not yet definite.

3 Admissible Experiments and Simulated Stochastic Process

With the aim of defining the adjustment curve for the stochastic process we assume has generated the data, we simulated experiments in some sense *similar* to the original one. More precisely, we required that each experiment consists in L curves, that the i-curve starts from $x^i(0)$ $(i = 1, \ldots, L)$ and that the frequency distribution of the values at T in the simulated experiment is close to the frequency distribution of the values at T of the real data (in terms of minimum and maximum of these data and of the χ^2 index). Experiments satisfying such conditions have been realized by using special RMCs. A Multiplicative Cascade is a single process that fragments a set into smaller and smaller components according to a fixed rule, and at the same time fragments the *measure* of components by another rule. The notion of multiplicative cascade was introduced in the statistical theory of turbulence by Kolmogorov (1941). Random cascade models have been used as models for a wide variety of other natural phenomena such as rainfall (e.g., see Gupta and Waymire 1993), internet packet traffic (e.g., see Resnick et al. 2003), market price (e.g., see Mandelbrot 1998). Recently statistical estimation theory for random cascade models has been investigated by Ossiander and Waymire (2000, 2002). We defined a RMC model generating recursively a multifractal measure μ on the family of all dyadic subintervals of the unit interval $[0, 1]$, and depending on a number of real positive parameters and constants obtained from the data (FD and R). Amongst these last constants are very important the ratio q^0 between the number of good realizations of the real process – that is the number of those curves whose outcome was good at time T – and the totality of such curves and p the number of steps of the multiplicative cascade, that was in our application $p = 41$. Each single launch of the RMC, truncated at the step p generates a *proof* of length $p + 1$, i.e., a single line of data which simulates a single row of the matrix FD. A set E_p of L proofs of length $p + 1$ satisfying previously outlined conditions is called an *Admissible Experiment* of size L and length $p + 1$ and can be rearranged in a matrix SFD of *Simulated Functional Data*. We denote by $S(E_p)$ the set of L piecewise linear interpolant the data in each single row of SFD. We define the stochastic process X as the set $X = \bigcup_{E_p \in \mathscr{E}_{\eta,\theta}} S(E_p)$ where by $\mathscr{E}_{\eta,\theta}$ we denote the set of all admissible experiments E_p of size L and length $p + 1$. The indexes η, θ are positive real numbers which provide a measure of the closeness of the simulated experiment to the real data. Further, if the value $X(T)$ (see Sect. 2) is unknown, we compute the middle point c of the interval which separates the *good* values from the *bad* values in the last column of the matrix FD and we set $X(T) = c$. A proof (curve) is declared good if its value at the final time T is greater or equal to $X(T)$; otherwise the proof is declared bad. Therefore the set of *good* values and that of *bad* values in the last column of the matrix SFD are separated and we define the binary response associated to X by $Y : X \to \{bad, good\}$, $Y(x) = Y_{E_p}(x)$ where the vector Y_{E_p} is the binary response associated to E_p. By analogy with the case of real data we introduce the adjustment curve $\gamma_{a,E_p} : [0, p] \to [0, 1]$ for the binary outcome Y_{E_p} of the

admissible experiment $E_p \in \mathscr{E}_{\eta,\theta}$. By the change of variable $\tau = \frac{p}{T} \cdot t$ $(t \in [0,T])$ we get, for every $E_p \in \mathscr{E}_{\eta,\theta}$, $\gamma_{a,E_p}(t) = \gamma_{a,E_p}(\frac{p}{T} \cdot t)(t \in [0,T])$. We note that the set $\{\gamma_{a,E_p} : E_p \in \mathscr{E}_{\eta,\theta}\} = \{\gamma_1, \gamma_2, \ldots, \gamma_N\}$ is finite. We now consider the random experiment "*obtain an admissible experiment* E_p" whose sample space is the infinite set $\mathscr{E}_{\eta,\theta}$. We set $\mathscr{E}^i_{\eta,\theta} = \{E_p \in \mathscr{E}_{\eta,\theta} : \gamma_{a,E_p} = \gamma_i\}$, $i = 1, \ldots, N$. Let ν_i be the frequencies of the curves γ_i, $i = 1, \ldots, N$: we define the *adjustment curve* $\gamma_a : [0,T] \to [0,1]$ for the binary response Y of the process X as the function $\gamma_a(t) = \sum_{i=1}^{N} \nu_i \gamma_i(t), t \in [0,T]$. In practice, given a couple $(X,Y) \in \mathscr{S} \times \mathscr{R}$ we can choose a tolerance $\epsilon > 0$ such that, if $E^1_p = (x^1_1, \ldots, x^1_L)$, $E^2_p = (x^2_1, \ldots, x^2_L)$ are two admissible experiments such that $\max_{i=1}^{L} \|x^1_i - x^2_i\|_\infty \leq \epsilon$ (here $\|\cdot\|_\infty$ denotes the usual sup-norm) then E^1_p, E^2_p can be considered indistinguishable. Therefore, X becomes a process with a discrete number of realizations and thus we can assume that for $i = 1, \ldots, N$, $\nu_i = \lim_{n\to\infty} \nu^n_i$, where ν^n_i is the relative frequency of γ_i observed on a sample $(\gamma_1, \ldots, \gamma_n)$ of size n. We set $\gamma^n_a = \sum_{i=1}^{n} \nu^n_i \gamma_i$ $(n = 1, 2, \ldots)$. The sequence $\{\gamma^n_a\}$ converges to γ_a on $[0,T]$ and the variance $Var(\gamma_a)$ of the random variable γ_a is less than or equal to two. Consequently the classical Monte Carlo method can be used to produce approximations of γ_a with the needed precision. The RMC Model synthetically described in this section and its computational aspects are fully detailed in Costanzo et al. (2010, 2009).

4 Application to an Industrial Kneading Process

We present an application of our method to a real industrial process; namely we will show how our model can be used to monitor and predict the quality of the outcome in an industrial kneading process. We will use a sample of data provided by Danone Vitapole Research Department (France). In kneading data from Danone, for a given flour, the resistance of dough is recorded during the first 480 s of the process. There are 136 different flours and thus 136 different curves or trajectories. Each one is obtained by Danone as a mean curve of a number of replications of the kneading process for each different flour. Each curve is observed in 240 equispaced time points (the same for all flours) of the interval time $[0, 480]$. Depending on its quality, after kneading, the dough is processed to obtain cookies. For each flour the quality of the dough can be *bad* or *good*. The sample we considered contains 44 *bad* and 62 *good* observations; it also contained 30 *undecided* observations, but these were discarded from the analysis. In Fig. 1 gray curves (black curves) are those corresponding to the cookies that were considered good quality (bad quality) at the end of the kneading process. Observe how the achieved dough density in $T = 480$

was the end quality variable mostly affecting the final quality of the cookies.[1] In fact, gray curves are located mainly in the high part of the graph (above the black curves), while black curves are located mainly in the low part of the graph (below the gray curves). However, in order to introduce the adjustment curve, we require that, with respect to the end values of the process, there is a clear separation between bad and good curves, that is R must depend only on the values at the time T of the real process (see Sect. 2). To meet such a condition we introduced (Costanzo et al. 2009) the concept of $\epsilon - (m, n)$ separability for two sets, which allows us to *clean* the data so that the ratio q^0 varies less than ϵ. In practice, by means of the $\epsilon - (m, n)$ separability, we find the minimum number of bad curves and/or good curves that can be discarded in a way that the ratio q^0 is kept within a preset error ϵ. For $\epsilon = 0.05$ we discarded from our analysis eight good curves and six bad curves; the remaining 54 good curves and 38 bad curves are separated in $T = 480$ at the dough resistance value $c = X(T) = 505$. In Fig. 2 we show one admissible experiment E_p obtained by the method outlined in Sect. 3. Observe that in obtaining bad/good trajectories in our cascade model we utilized as parameters set by the data q^0 and $1-q^0$ respectively. However, owing to the randomness of our model the numbers of bad and good trajectories are not necessarily in the same proportion as in the original data. We remark that in applying the Monte Carlo Method in order to obtain the adjustment curve γ_a with an error less than 10^{-1} and probability greater than 90% we need to perform $n = 4{,}000$ admissible experiments. In Fig. 3 we depicted the adjustment curve γ_a for the binary response Y of the stochastic process X related to the Danone data. This curve has been computed on the basis of $n = 1{,}000$ admissible experiment E_p, obtained requiring a value of the χ^2 index less than or equal

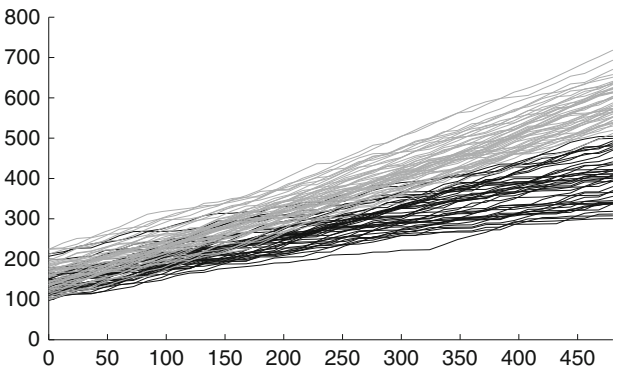


Fig. 2 An admissible experiment E_p

[1] Actually, decision good/bad/undecidable about the quality of the dough was in the Danone case based on an taster judgement which presumably also took in account other quality variables such as aroma, etc. This may be the reason why some of the doughs were classified as undecidables. In our analysis we decided to consider the density of the dough as the only end quality variable.

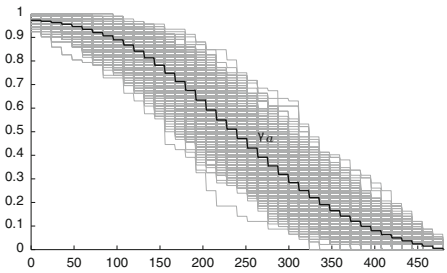


Fig. 3 The adjustment curves of the process and of 1,000 admissible experiments

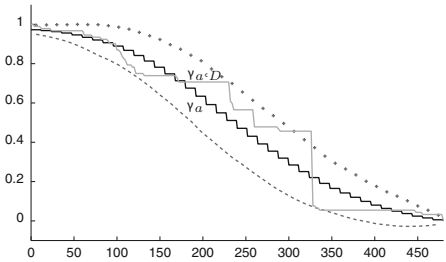


Fig. 4 The adjustment curves of the process and of Danone's data

to seven. In the same figure we also depicted in gray the adjustment curves of these admissible experiments. Observe that, for each $t \in [0, T]$, the standard deviation is not great than $0.09 \approx 10^{-1}$. The intervals of one standard deviation from the mean curve value comprise a frequency of adjustment curves of E_p's admissible experiments which range from a minimum of about 65% – in the time interval between $t = 150$ and $t = 320$ about – to a maximum of about 87%; while in the intervals of two standard deviations, the frequency range is between 92 and 97%. These latter intervals are illustrated in Fig. 4 by the plus and minus signs and they comprise the adjustment curve $\gamma_{a,D}$ of the real Danone data. In this figure we can observe that the simulated process prediction curve gives the same results as the real data at times near to $t = 186$, which was the same time $t^* < T$ we determined in Costanzo et al. (2006) on the whole interval $[0, 480]$ (the average test error rate was of about 0.112 for an average $\overline{AUC}(T) = 0.746$). However, after such time and until time $t = 326$ the $\gamma_{a,D}$ gives an adjustment higher than γ_a, which denotes instability for such data since from 65 to 42% of the curves are not yet definitively bad or good (see the bold black and gray curves in Fig. 1). Let us remark that the mean of the absolute differences between the two curves is 0.06 over the whole time interval, while its value is 0.112 in the time interval $[186, 326]$. Starting from time $t = 330$ γ_a is over $\gamma_{a,D}$ instead. Consider then for greater clarity to use γ_a to predict the bad outcomes and $1 - \gamma_a$ to predict for good outcomes: an adjustment value of about $\gamma_a = 0.20$ implies, that bad outcomes after such time have low probability (≤ 0.20) of adjustment before the end of the process and they could be discarded or the process modified; while, at the same time, a good outcome has high probability ($1 - \gamma_a = 0.80$) to remain the same until the end of the process. As we pointed out in the introduction, the adjustment curve γ_a is an important monitoring tool of a real process: it represents a prediction function by means of which, as in the case

of cookies, the quality of product at the final time T, can be anticipated for each given $t \in [0, T]$ with an increasing probability. Since in real processes the interest is mainly in preventing bad outcomes, the term *adjustment* means that, as long as the values of such curve are high, it is not necessary to intervene or stop the process, because with any performance which it seems will result in a bad outcome, there is a high probability of (self)adjustment before the end of the process itself. Further, the same definition of adjustment curve allows for decisions at each time $t \in [0, T]$ on the *quality* of the outcome of each single realization of the process at that time. In fact, as shown in Sect. 2, since for real data this curve is obtained from the two sets (1) and (2), we consider in the same way for each admissible experiment E_p the piecewise linear functions b_{E_p} e g_{E_p} whose node-sets are defined as in the previous equations. A single realization (curve) of the process is then good (bad) in t if at this time its value is greater(lower) or equal to the mean value of the b_{E_p}'s (g_{E_p}'s). Otherwise the status of the realization at this time is not yet determinable and to simplify matters we can cautiously decide to treat it as it were a bad realization (if a process realization has not a defined status until the last time $T - 1$, it will have at time T one-half of probability to result in a bad realization and this could be rather expensive in terms of cost of the process). Consider then for more clearness to use γ_a to predict on the bad outcomes and $1 - \gamma_a$ to predict for good outcomes.The adjustment curve gives at each given time t the probability that a bad realization will change its status before the end of the process; while its complementary gives the probability that a good realization maintains its status until the end of the process.

5 Conclusion and Open Issues

Some open issues involve further research to improve the choice of the parameters of the RMC Model so that it is more flexible to adapt to several situations and time dynamics of different observed curves; study of the sampling distribution of our method; embedding of our results in a cost (or in general a loss) function so that an optimal in some sense, prediction time is elicited; extension of the model to the case of polytomous response variables.

Acknowledgements Thanks are due for their support to Food Science & Engineering Interdepartmental Center of University of Calabria and to L.I.P.A.C., Calabrian Laboratory of Food Process Engineering (Regione Calabria APQ-Ricerca Scientifica e Innovazione Tecnologica I atto integrativo, Azione 2 laboratori pubblici di ricerca mission oriented interfiliera).

References

Box, G., & Kramer, T. (1992). Statistical process monitoring and feedback adjustment-A discussion. *Technometrics*, *34*(3), 251–267.

Costanzo, G. D., Preda, C., & Saporta, G. (2006). Anticipated prediction in discriminant analysis on functional data for binary response. In A. Rizzi & M. Vichi (Eds.), *COMPSTAT'2006 Proceedings* (pp. 821–828). Heidelberg: Springer.

Costanzo, G. D., Dell'Accio F., & Trombetta, G. (2009). Adjustment curves for binary responses associated to stochastic processes. *Dipartimento di Economia e Statistica*, Working Paper no 17, Anno 2009.

Costanzo, G. D., De Bartolo, S., Dell'Accio F., & Trombetta, G. (2010). Using observed functional data to simulate a stochastic process via a random multiplicative cascade model. In Y. Le Chevalier & G. Saporta (Eds.), *COMPSTAT2010 Proceedings*, pp. 453–460. Heidelberg: Springer.

Escabias, A. M., Aguilera, A. M., & Valderrama, M. J. (2007). Functional PLS logit regression model. *Computational Statistics and Data Analysis*, *51*, 4891–4902.

Gupta, V. K., & Waymire, E. (1993). A statistical analysis of mesoscale rainfall as a random cascade. *Journal of Applied Meteorology*, *32*, 251–267.

James, G. (2002). Generalized linear models with functional predictor variables. *Journal of the Royal Statistical Society Series B*, *64*, 411–432.

Kesavan, P., Lee, J. H., Saucedo, V., & Krishnagopalan, G. A. (2000). Partial least squares (PLS) based monitoring and control of batch digesters. *Journal of Process Control*, *10*, 229–236.

Kolmogorov, A. N. (1941). The local structure of turbulence in incompressible viscous fluid for very large Reynolds number. *Doklady Akademii nauk SSSR*, *30*, 9–13.

Mandelbrot, B. (1998). *Fractals and scaling in finance: Discontinuity, concentration, risk*. New York: Springer-Verlag.

Ossiander, M., & Waymire, C. E. (2000). Statistical estimation for multiplicative cascades. *The Annals od Statistics*, *28*(6), 1533–1560.

Ossiander, M., & Waymire, C. E. (2002). On estimation theory for multiplicative cascades. *Sankhyā, Series A*, *64*, 323–343.

Preda C., & Saporta, G. (2005). PLS regression on a stochastic process. *Computational Satistics and Data Analysis*, *48*, 149–158.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis*. Springer series in statistics. New York: Springer-Verlag.

Resnick, S., Samorodnitsky, G., Gilbert, A., & Willinger, W. (2003). Wavelet analysis of conservative cascades. *Bernoulli*, *9*, 97–135.

Saporta, G., Costanzo, G. D., & Preda, C. (2007). Linear methods for regression and classification with functional data. In *IASC-ARS'07 Proceedings, Special Conf., Seoul, 2007* (ref. CEDRIC 1234).

Dealing with FDA Estimation Methods

Tonio Di Battista, Stefano A. Gattone, and Angela De Sanctis

Abstract In many different research fields, such as medicine, physics, economics, etc., the evaluation of real phenomena observed at each statistical unit is described by a curve or an assigned function. In this framework, a suitable statistical approach is Functional Data Analysis based on the use of basis functions. An alternative method, using Functional Analysis tools, is considered in order to estimate functional statistics. Assuming a parametric family of functional data, the problem of computing summary statistics of the same parametric form when the set of all functions having that parametric form does not constitute a linear space is investigated. The central idea is to make statistics on the parameters instead of on the functions themselves.

1 Introduction

Recently, Functional data analysis (FDA) has become an interesting research topic for statisticians. See for example Ferraty and Vieu (2006) and Ramsay and Silverman (2007) and reference therein. In many different fields, data come to us through a process or a model defined by a curve or a function. For example, in psychophisiological research, in order to study the electro dermal activity of an individual, the Galvanic Skin Response (GSR signal) can be recorded and represented by a continuous trajectory which can be studied by means of the tools of FDA (Di Battista et al. 2007). We want to deal with circumstances where functional data are at hand and the function is known in its closed form. In particular, we consider a parametric family of functional data focusing on parameters estimation of the function. For example, Cobb-Douglas production functions are frequently used in economics in order to study the relationship between input factors and the level of production. This family of functions takes on the form $y = f(K, L) = L^{\alpha} K^{\beta}$, where L is one factor of production (often labour) and K is a second factor of production (often capital) and α and β are positive parameters with $\alpha + \beta = 1$. In biology, growth functions are used to describe growth processes (Vieira and Hoffmann 1977). For example, the logistic growth function $Z = a/[1 + \exp\{-(b + ct)\}]$ where a, b and c are parameters,

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_40,

$a > 0$ and $c > 0$, and the Gompertz growth function $Z = \exp\left(a - bc^t\right)$ where a, b and c are parameters, $b > 0$ and $0 < c < 1$. The aims of FDA are fundamentally the same as those of any area of statistics, i.e., to investigate essential aspects such as the mean and the variability function of the functional data. Moreover, one could be interested in studying the rate of change or derivatives of the curves. However, since functional data are often observed as a sequence of point data, then the function denoted by $y = x(t)$ reduces to a record of discrete observations that we can label by the n pairs $\left(t_j, y_j\right)$ where y_j is the value of the function computed at the point t_j. A first step in FDA is to convert the values $y_{i1}, y_{i2}, \ldots y_{in}$ for each unit $i = 1, 2, \ldots, m$ to a functional form computable at any desired point t. To this purpose, the use of basis functions ensures a good fit in a large spectrum of cases. The statistics are simply those evaluated at the functions pointwise across replications.

It is well known that the sample mean $\bar{x}(t) = \frac{1}{m}\sum_{i=1}^{m} x_i(t)$ is a good estimate of the mean if the functional data are assumed to belong to L^2. If we do not need of a scalar product and then of an orthogonality notion, we can consider every L^p space, $p > 1$, with the usual norm (Rudin 2006). In general the functional data constitute a space which is not a linear subspace of L^p. For example, let $y_1 = A_1 L^{\alpha_1}$ and $y_2 = A_2 L^{\alpha_2}$ be Cobb-Douglas functions in which for simplicity the production factors A_1 and A_2 are assumed constant. The mean function is $\bar{y} = \frac{A_1 L^{\alpha_1} + A_2 L^{\alpha_2}}{2}$ which is not a Cobb-Douglas function and its parameter does not represent the well known labour elasticity which is crucial to evaluate the effect of labour on the production factor. In general, the results of this approach may not belong to a function with the same closed form of the converted data so that erroneous interpretations of the final functional statistic could be given.

In this communication we want to emphasize a new approach which is focused on the true functional form generating the data. First of all, we introduce a suitable interpolation method (Sung Joon 2005) that allows us to estimate the function that is suspected to produce the functional datum for each replication unit. Starting from the functional data we propose an explicit estimation method. The objective is to obtain functional statistics that belong to the family of functions or curves suspected to generate the phenomenon under study. In the case of a parametric family of functional data, we use the parameter space in order to transport the mean of the parameters to the functional space. Assuming a monotonic dependence from parameters we can obtain suitable properties for the functional mean. At illustrative purpose, two small simulation studies are presented in order to explore the behaviour of the approach proposed.

2 Orthogonal Fitting Curve and Function

Generally, functional data are recorded discretely as a vector of points for each replication unit. Thus, as a first step we need to convert the data points to a curve or a function. Methods such as OLS and/or GLS do not ensure the interpolation of a wide class of curves or functions. A more general method is given by the

Least Squares Orthogonal Distance Fitting of Curves (ODF) (Sung Joon 2005). The goal of the ODF is the determination of the model parameters which minimize the square sum of the minimum distances between the given points $\{\mathbf{Y}_j\}_{j=1}^n$ and the closed functional form belonging to the family of curves or function $\{f(\boldsymbol{\theta}, t)\}$ with $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_p\}$. In ODF the corresponding points $\left\{\mathbf{Y}_j^*\right\}_{j=1}^n$ on a fitted curve are constrained to being membership points of a curve/surface in space. So, given the explicit form $f(\boldsymbol{\theta}; t)$ such that $\mathbf{Y}^* - f(\boldsymbol{\theta}; t) = 0$, the problem leads to minimize a given cost function. Two performances indices are introduced which represent in two different ways the square sum of the weighted distances between the given points and the functional form $f(\boldsymbol{\theta}; t)$: the performances index $\sigma_0^2 = \|P(\mathbf{Y} - \mathbf{Y}^*)\|^2 = (\mathbf{Y} - \mathbf{Y}^*)^T \mathbf{P}^T \mathbf{P}(\mathbf{Y} - \mathbf{Y}^*)$ in coordinates based view or $\sigma_0^2 = \|\mathbf{Pd}\|^2 = \mathbf{d}^T \mathbf{P}^T \mathbf{Pd}$ in distance based view, where $\mathbf{P}^T\mathbf{P}$ is a weighting matrix or error covariance matrix (positive definite), $\mathbf{Y}^* = \left\{\mathbf{Y}_j^*\right\}_{j=1}^n$ is a coordinate column vector of the minimum distance points on the functional form from each given point $\{\mathbf{Y}_j\}_{j=1}^n$, $\mathbf{d} = (d_1, d_2, \ldots, d_n)^T$ is the distance column vector with $d_j = \left\|\mathbf{Y}_j - \mathbf{Y}_j^*\right\| = \sqrt{\left(\mathbf{Y}_j - \mathbf{Y}_j^*\right)^T \left(\mathbf{Y}_j - \mathbf{Y}_j^*\right)}$. Using the Gauss Newton method, it is possible to estimate the model parameters $\boldsymbol{\theta}$ and the minimum distance points $\left\{\mathbf{Y}_j^*\right\}_{j=1}^n$ with a variable separation method in a nested iteration scheme as follows

$$\min_{\boldsymbol{\theta} \in R^p} \min_{\left\{\mathbf{Y}_j^*\right\}_{j=1}^n \in Z} \sigma_0^2 \left(\{\mathbf{Y}_i^*(\boldsymbol{\theta})\}_{j=1}^n\right) \tag{1}$$

whith $Z = \left\{\mathbf{Y} \in R^n : \mathbf{Y} - f(\boldsymbol{\theta}; t) = 0, \boldsymbol{\theta} \in R^p, t \in R^k\right\}$.

3 Direct FDA Estimation Methods

Let S be a family of functions with p real parameters that is $S = \{f_{\boldsymbol{\theta}}\}$ with $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots \theta_p) \in \Theta$. In an economic setting, S could be the family of Cobb-Douglas production functions, i.e., $f_{\alpha,\beta}(K, L) = K^\alpha L^\beta$ with $\alpha > 0$, $\beta > 0$ and $\alpha + \beta = 1$. Starting from m functional data belonging to S, $f_{\boldsymbol{\theta}_1}, f_{\boldsymbol{\theta}_2}, \ldots, f_{\boldsymbol{\theta}_m}$, the objective is to find an element of S said functional statistic denoted with $f_{\hat{\boldsymbol{\theta}}} = H\left(f_{\boldsymbol{\theta}_1}, f_{\boldsymbol{\theta}_2}, \ldots, f_{\boldsymbol{\theta}_m}\right)$.

3.1 The Functional Mean

In the following we assume that functional data constitute a subspace S of some L^p space, $p > 0$, with the usual norm (Rudin 2006). We consider first the functional mean of the functions $f_{\theta_1}, f_{\theta_2}, \ldots, f_{\theta_m}$. When S is a vectorial subspace, then we can express the functional mean as the sample mean $f_{\hat{\theta}} = \frac{f_{\theta_1} + f_{\theta_2} + \ldots + f_{\theta_m}}{m}$.

Because S is closed with respect to linear combinations, we have that $f_{\hat{\theta}} \in S$. In this setting a straightforward property is that the integral of the functional mean is the mean of the integrals of each functional datum. For example, let S be the family of functions of the following form $f_\alpha = \alpha g\,(x)$, then

$$f_{\hat{\alpha}} = \frac{\sum_{i=1}^{m} f_{\alpha_i}(x)}{m} = \frac{\sum_{i=1}^{m} \alpha_i\, g\,(x)}{m} = \frac{\sum_{i=1}^{m} \alpha_i}{m} g\,(x)\,. \tag{2}$$

This proves that $f_{\hat{\alpha}}\,(x)$ is an element of S and its parameter is the mean of the parameters $\alpha_1, \alpha_2, \ldots, \alpha_m$. At the same time it is easy to prove that if S is not a vectorial space then this functional statistic doesn't necessarily lead to an element belonging to S. We go along in two ways. The first one is to verify if there is an element in S that has got as integral the mean of the integrals of the functional data. For instance, let S be the family of functions $f_\alpha\,(x) = x^\alpha$, with $0 < \alpha < 1$ and domain the closed interval $[0, 1]$. If $m = 2$, then $\int_0^1 x^\alpha dx = \frac{\int_0^1 x^{\alpha_1} dx + \int_0^1 x^{\alpha_2} dx}{2}$, that is $\frac{1}{\alpha+1} = \frac{\frac{1}{\alpha_1+1} + \frac{1}{\alpha_2+1}}{2}$ which admits a unique solution. For example if we have got two functions with parameters $\alpha_1 = \frac{1}{2}$ and $\alpha_2 = \frac{1}{3}$ then $\hat{\alpha} = \frac{7}{17}$. Unfortunately, in general the solution may not exist in the real field and/or it is not unique and it would be necessary to introduce some constraints on the parameters not easy to interpret.

A second way to solve the problem without ambiguity is the following. We assume that every functional datum f_θ is univocally determined by the parameter θ or equivalently there is a biunivocal correspondence between S and the parameter space Θ. Then, a functional statistic for the space of the functional data can be obtained through a statistic in the parameter space. In the case of a parametric family of functional data, we use the parameter space in order to transport the statistics in Θ to S. Let the functional data be $f_{\theta_1}, f_{\theta_2}, \ldots, f_{\theta_m}$, then a functional statistic for the set of the functional data is given by a suitable statistic of the parameters $\theta_1, \theta_2, \ldots, \theta_m$ say $\hat{\theta} = K\,(\theta_1, \theta_2, \ldots, \theta_m)$. The functional statistic will be the element of S that has got as parameter the statistic $\hat{\theta}$, following the scheme:

$$\begin{array}{rl} \theta_i & \leftarrow f_{\theta_i} \\ \downarrow & \qquad i = 1, 2, \ldots, m \\ \hat{\theta} = K\,(\theta_i) & \rightarrow f_{\hat{\theta}}. \end{array} \tag{3}$$

A possible way of defining the function K is the analogy criterion. If we want to estimate the functional mean or median then the function K would be the mean or the median of the parameters. Obviously, other ways of defining the function K are possible. The advantage in this case is that we can require for the functional mean and variability the same properties of the mean and variance of the parameters. In particular, for the functional mean, we can assume that the functions are linked to each parameter by a monotonic dependence. For example, if we have only a parameter α, we can suppose $\alpha_1 \leq \alpha_2 \Rightarrow f_{\alpha_1}\,(x) \leq f_{\alpha_2}\,(x)$ or $f_{\alpha_1}\,(x) \geq f_{\alpha_2}\,(x)\ \forall x$. In such a case, for the mean parameter $\hat{\alpha}$, we obtain $f_{\alpha_1}\,(x) \leq f_{\hat{\alpha}}\,(x) \leq f_{\alpha_2}\,(x)\ \forall x$. Moreover this property ensures also that $\int f_{\alpha_1}\,(x)dx \leq \int f_{\hat{\alpha}}\,(x)dx \leq \int f_{\alpha_2}\,(x)dx$. It

is easy to verify that monotonic decreasing dependence is verified by the family $S = f_\alpha(x) = x^\alpha$ with $0 < \alpha < 1$ and $x \in [0, 1]$.

3.2 Functional Variability

In order to study the functional variability we first introduce the functional quantity $v_i^r(t) = \left| f_{\theta_i}(t) - f_{\hat{\theta}}(t) \right|^r$ which is the r-th order algebraic deviation between the functional observed data f_{θ_i} and the functional statistics $f_{\hat{\theta}}$. Then the functional variability can be measured pointwise by the r-th order functional moment

$$V^r(t) = \frac{1}{m} \sum_{i=1}^{m} v_i^r(t). \tag{4}$$

The function $V^r(t)$ has the following properties:

- if $f_{\theta_i}(t) = f_{\hat{\theta}}(t)$ for $i = 1, 2, \ldots, m$ and $\forall t$, than $V^r(t) = 0$;
- defining the L^p norm of a function as $\| f_\theta(t) \|_{L^p} = \int |f_\theta(t)|^p \, dt$ then we have that

$$\left\{ \left\| f_{\theta_i} - f_{\hat{\theta}} \right\|_{L^p} \to 0 \right\} \Rightarrow \left\{ f_{\theta_i} \overrightarrow{a.e.} \; f_{\hat{\theta}} \Leftrightarrow v_i^r \overrightarrow{a.e.} \; 0 \; \forall i = 1, 2, \ldots, m \Leftrightarrow V^r \overrightarrow{a.e.} \; 0 \right\}.$$

We remark that, if the function f_θ in S is expandable in Taylor's series, that is

$$f_\theta(t) = \sum_{k=0}^{\infty} \frac{f_\theta^k(a)}{k!} (t - a)^k \tag{5}$$

where a is a fixed point of an open domain and $f_\theta^k(a)$ is the k-th derivative of the function f_θ computed at point a, an approximation of the functional variability can be obtained by Taylor's polynomials s_{θ_i} of f_{θ_i} and $s_{\hat{\theta}_i}$ of $f_{\hat{\theta}_i}$ respectively:

$$\frac{1}{m} \sum_{i=1}^{m} \left| s_{\theta_i}(t) - s_{\hat{\theta}_i}(t) \right|^p. \tag{6}$$

This fact is useful from a computation point of view. In order to give some insights to the approach proposed in the next section two small simulation studies are proposed.

4 A Simulation Study

We conduct two small simulation studies in order to evaluate the estimation method proposed for the functional statistic $f_{\hat{\theta}} = H\left(f_{\theta_1}, f_{\theta_2}, \ldots, f_{\theta_m}\right)$ equal to the functional mean.

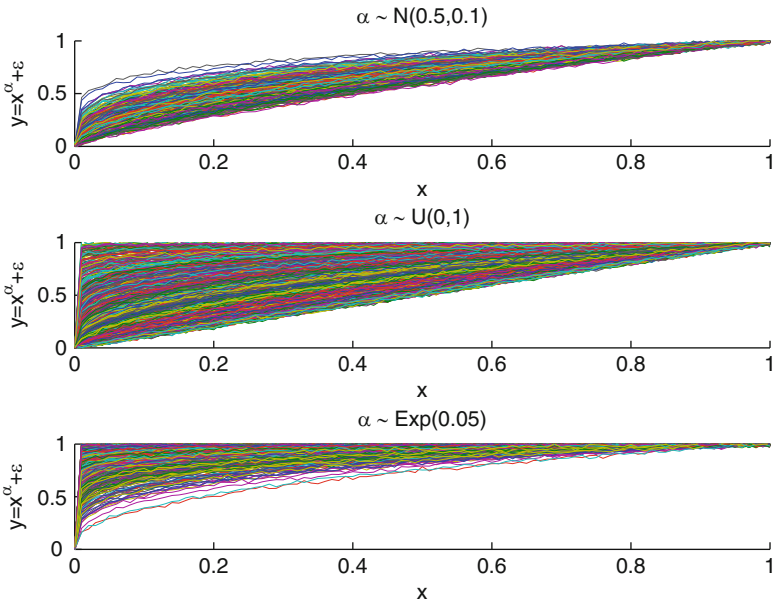


Fig. 1 Functional populations $S = \{f_\theta\} = x^\alpha + \epsilon$ with three different space parameter θ and $\epsilon \sim N(0, 0.01)$

4.1 Power Functions

We suppose that the observations are contaminated with some error so that the resulting family $S = \{f_\theta\}$ of functions is defined as $S = \{x^\alpha\} + \epsilon$ with $\theta = \alpha \in R^1$ with $0 < \alpha < 1$ and $0 \leq x \leq 1$. We simulate different populations by assigning to α different distributions such as the truncated Normal, the Uniform and the truncated Exponential with different parameters and to ϵ a white noise with standard error equal to 0.01. At illustrative purpose in Fig. 1 there are three populations for $\alpha \sim N(\mu = 0.5, \sigma = 0.1)$, $\alpha \sim U(0, 1)$ and $\alpha \sim Exp(0.05)$. Values of α outside the interval $(0, 1)$ were discarded.

In order to evaluate the estimation method proposed in Sect. 3, we sample from each population $J = 5{,}000$ samples for various sample sizes m. As the functions are observed with error we first need to apply the ODF method of Sect. 2 to estimate the function parameter α for each function. Once for each sample the estimates $\theta_1, \theta_2, \ldots, \theta_m$ are available, the scheme detailed in (3) can be applied in order to obtain the functional mean statistic of the sample. In Fig. 2 we show the results for a sample size of $m = 10$. In particular, for each population, the functional mean statistic together with the estimated standard error are plotted.

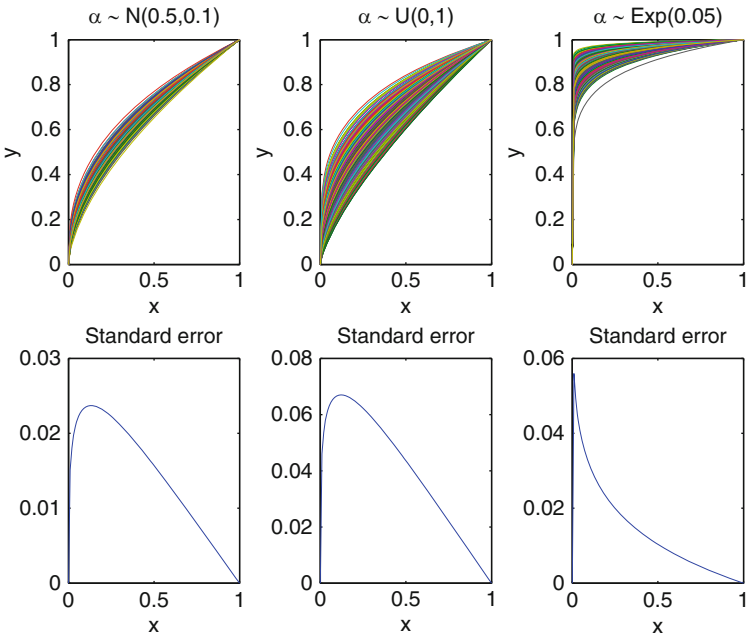


Fig. 2 J = 5,000 Functional mean statistics for a sample size $m = 10$

4.2 Functional Diversity Profiles

At illustrative purpose, we present an ecological application of the estimation method proposed. Suppose to have a biological population made up of p species where we are able to observe the relative abundance vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ in which the generic θ_j represents the relative abundance of the j-th species. One of the most remarkable aspects in environmental studies is the evaluation of ecological diversity. The most frequently used diversity indexes may be expressed as a function $f_{\boldsymbol{\theta}}$ of the relative abundance vector. Patil and Taillie (1982) proposed to measure diversity by means of the β-diversity profiles defined as

$$\Delta = f_{\boldsymbol{\theta}}(\beta) = \frac{1 - \sum_{j=1}^{p} \theta_j^{\beta+1}}{\beta}. \tag{7}$$

β-diversity profiles are non-negative and convex curves. In order to apply functional linear models on diversity profiles, Gattone and Di Battista (2009) applied a transformation which can be constrained to be non-negative and convex. In the FDA context, it is convenient considering the β-diversity profile as a parametric function computable for any desired argument value of $\beta \in [-1, 1] \setminus \{0\}$. The space parameter is multivariate and given by $\boldsymbol{\theta}$. In order to evaluate the estimation method proposed in Sect. 3, we simulate different biological populations by assigning to each component of $\boldsymbol{\theta}$ different distributions such as the Uniform, the Poisson and

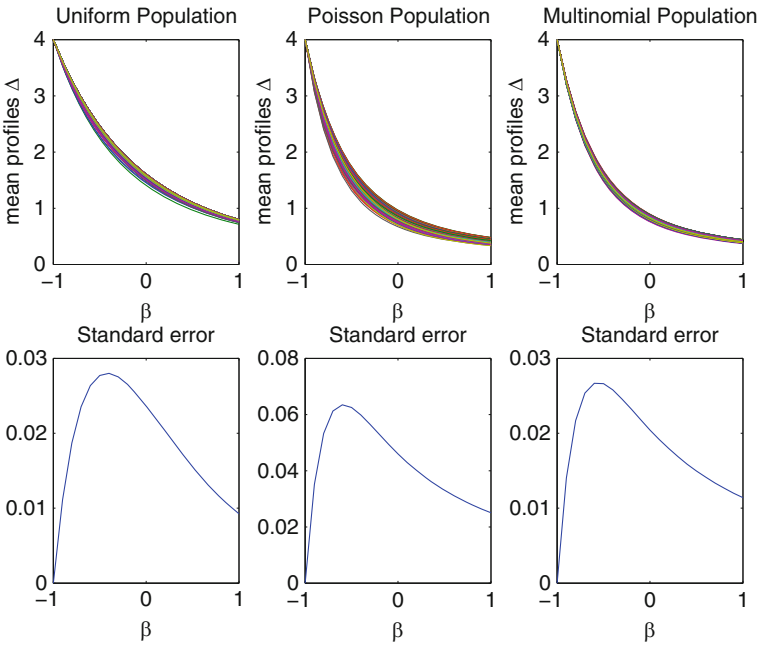


Fig. 3 J = 5,000. Functional mean diversity profiles $\Delta = f_{\hat{\boldsymbol{\theta}}} = \frac{1-\sum_{j=1}^{p} \theta_j^{\beta+1}}{\beta}$ and standard error for a sample size $m = 5$

the multinomial distribution. From each population we sample $J = 5,000$ samples with different sample sizes. The function Δ in (7) is observed without error so that we do not need to apply the ODF method of Sect. 2. For each sample of size m we can evaluate the estimates $\hat{\boldsymbol{\theta}}$ from the observed $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_m$ and the scheme detailed in (3) can be applied in order to obtain the functional mean statistic $\Delta = f_{\hat{\boldsymbol{\theta}}}$. In Fig. 3 we show the results for three populations with $p = 5$ species with different level of diversity. From each population we randomly choose samples of size $m = 5$. The parameters of the Poisson and the Multinomial distributions are $\lambda = 100 * [0.55, 0.19, 0.13, 0.07, 0.06]$ and $[0.55, 0.19, 0.13, 0.07, 0.06]$, respectively. For each population, the functional mean statistic together with the estimated standard error are plotted. As desired, all the functional statistics result to be non-negative and convex. Furthermore, even though monotonic dependence from the parameters is not verified with diversity profiles, the functional mean satisfies the internality property in all the simulation runs.

References

Di Battista, T. Gattone S.A., & Valentini, P. (2007). Functional Data Analysis of GSR signal, Proceedings *S.Co. 2007: Complex Models and Computational Intensive Methods for Estimation and Prediction*, CLEUP Editor, Venice, 169–174.

Ferraty, F., & Vieu, P. (2006). *Nonparametric functional data analysis: Theory and practice*. New York: Springer-Verlag.

Gattone, S. A., & Di Battista, T. (2009). A functional approach to diversity profiles. *Journal of the Royal Statistical Society, Series C*, *58*, 267–284.

Patil, G. P., & Taillie, C. (1982). Diversity as a concept and its measurements. *Journal of the American Statistical Association*, *77*, 548–561.

Ramsay, J. O., & Silverman, B. W. (2007). *Functional data analysis*. New York: Springer.

Rudin, W. (2006). *Real and complex analysis*. McGraw-Hill, New York.

Sung Joon, A. (2005). *Least squares orthogonal distance fitting of curves and surfaces in space*. New York: Springer.

Vieira, S., & Hoffmann, R. (1977). Comparison of the logistic and the Gompertz growth functions considering additive and multiplicative error terms. *Applied Statistics*, *26*, 143–148.

Part IX
Computer Intensive Methods

Testing for Dependence in Mixed Effect Models for Multivariate Mixed Responses

Marco Alfó, Luciano Nieddu, and Donatella Vicari

Abstract In regression modelling for multivariate responses of mixed type, the association between outcomes may be modeled through dependent, outcome-specific, latent effects. Parametric specifications of this model already exist in the literature; in this paper, we focus on model parameter estimation in a Finite Mixture (FM) framework. A relevant issue arises when independence should be tested vs dependence. We review the performance of LRT and penalized likelihood criteria to assess the presence of dependence between outcome-specific random effects. The model behavior investigated through the analysis of simulated datasets shows that AIC and BIC are of little help to test for dependence, while bootstrapped LRT statistics performs well even with small sample sizes and limited number of bootstrap samples.

1 Introduction

Regression models for multivariate responses have raised great interest in the last few years, with a particular emphasis on multivariate counts modeling. Three main approaches have been proposed: convolution models, see e.g., Karlis and Meligkotsidou (2006), latent effect models, see e.g., Chib and Winkelmann (2001) and Alfó and Trovato (2004), and copula-based models, see e.g., Harry (1997), the former not being applicable when dealing with mixed-response models. We adopt the latent effect approach and define a set of conditional univariate models, linked by a common latent structure which accounts for both heterogeneity (in the univariate profiles) and dependence between responses.

Let us suppose we have recorded responses Y_{ij}, on $i = 1, \ldots, n$ individuals and $j = 1, \ldots, J$ outcomes, together with a set of m_j covariates $\mathbf{x}_{ij}^{\mathsf{T}} = (x_{ij1}, \ldots, x_{ijm_j})$. To describe association among outcomes, it is reasonable to assume that they share some common unobservable features. Let u_{ij}, $i = 1, \ldots, n$ $j = 1, \ldots, J$ denote a set of individual and outcome-specific random effects, accounting for marginal heterogeneity and dependence between outcomes. Conditional on the covariates

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_41,

and the random effects, the observed responses Y_{ij} are assumed to be independent exponential family random variables

$$Y_{ij}|u_{ij}, \mathbf{x}_{ij} \sim \text{EF}(\theta_{ij})$$

with canonical parameters $\theta_{ij} = m^{-1}\left[E(Y_{ij} \mid \mathbf{x}_{ij}, u_{ij})\right]$, where $m(\cdot)$ denotes the mean function, modeled as follows:

$$\theta_{ij} = \beta_{j0} + \sum_{l=1}^{m_j} x_{ijl}\beta_{jl} + u_{ij} \qquad i = 1, \ldots n; \quad j = 1, \ldots J \tag{1}$$

Responses of different types can be modeled using this approach as long as each distribution is a member (not necessarily the same) of the exponential family. In this context, $\beta_j = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{jm_j})$ is an outcome-specific vector of fixed regression parameters, while the random effects $\mathbf{u}_i = (u_{i1}, \ldots, u_{iJ})$ are assumed to be drawn from a known multivariate parametric distribution, say $\mathscr{G}(\cdot)$, with $E(\mathbf{u}_i) = \mathbf{0}$. Semiparametric multivariate alternatives with unspecified $\mathscr{G}(\cdot)$ are discussed in Alfó and Trovato (2004) for multivariate counts and in Alfó et al. (2011) for multivariate counts and a binary selection variable. Under the assumption of conditional independence, the likelihood function is:

$$L(\cdot) = \prod_{i=1}^{n}\left\{\int_{\mathscr{U}}\left[\prod_{j=1}^{J} f\left(y_{ij} \mid \mathbf{x}_{ij}, u_{ij}\right)\right] d\mathscr{G}(\mathbf{u}_i)\right\} \tag{2}$$

For Gaussian assumptions on $\mathbf{u}_i$, the marginal likelihood can not be written in closed form; to obtain ML estimates, we may choose among several alternatives. We may adopt numerical integration techniques based on standard or adaptive (spherical) Gaussian Quadrature (GQ, AGQ), for a comparison see Rabe-Hesket and Skrondal (2002) and Rabe-Hesket et al. (2005); however, the corresponding estimation algorithms can be cumbersome and show slow convergence when the number of outcomes increases. A further alternative is to rely on Monte Carlo or simulation-based techniques, see e.g., Chib and Winkelmann (2001) and Munkin and Trivedi (1999). The latter approach is inefficient for non optimal importance samplers, while the distribution of the random effects conditional on the observed data and the current parameter estimates can be quite difficult to sample from. Marginal maximization procedures using Gaussian quadrature or Monte Carlo approximations can be computationally intensive, as noted in different contexts by Crouch and Spiegelman (1990) and Gueorguieva (2001). From this perspective, a more appealing approach is to leave $\mathscr{G}(\cdot)$ unspecified, and rely on the theory of NPML, see e.g., Kiefer and Wolfovitz (1956), Laird (1978) and Heckman and Singer (1984). If the likelihood is bounded, it is maximized with respect to $\mathscr{G}(\cdot)$ by at least a discrete distribution $\mathscr{G}_K(\cdot)$ with at most $K \leq n$ support points. Let us suppose that $\mathscr{G}_K(\cdot)$

puts masses π_k on locations $\mathbf{u}_k = (u_{k1}, \ldots, u_{kJ})$, $k = 1, \ldots, K$. The resulting likelihood function is:

$$L(\cdot) = \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} \pi_k \left[\prod_{j=1}^{J} f\left(y_{ij} \mid \mathbf{x}_{ij}, u_{kj}\right) \right] \right\} \tag{3}$$

where $\pi_k = \Pr(\mathbf{u}_i = \mathbf{u}_k) = \Pr(u_{k1}, \ldots, u_{kJ}), k = 1, \ldots, K$.

2 Computational Details

The data vector is composed by an observable part, $\mathbf{y}_i$, and by an unobservable part, $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$ representing the membership vector. Since the $\mathbf{z}_i$ are unknown, an EM-type algorithm can be used. For fixed K, the following routine for parameter estimation can be used: we assume that $\mathbf{z}_i$ has a multinomial distribution with weights $\pi_k = \Pr(\mathbf{u}_i = \mathbf{u}_k), i = 1, \ldots, n, k = 1, \ldots, K$. Given the model assumptions, the log-likelihood for the complete data is:

$$\ell_c(\cdot) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left\{ \log(\pi_k) + \log(f_{ik}) \right\} = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left\{ \log(\pi_k) + \sum_{j=1}^{J} \log(f_{ijk}) \right\} \tag{4}$$

Within the E-step, the presence of missing data is handled by taking the expectation $Q(\cdot)$ of the log-likelihood for the complete data given the observed data $\mathbf{y}_i$ and the current ML parameter estimates, say $\hat{\boldsymbol{\theta}}^{(r)}$, i.e., z_{ik} is replaced by its conditional expectation $w_{ik}^{(r)}, i = 1, \ldots, n, k = 1, \ldots, K$

$$Q(\cdot) = \mathrm{E}_{\boldsymbol{\theta}^{(r)}} \left\{ \ell_c(\cdot) \,|\mathbf{y}_i \right\} = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(r)} \left\{ \log(\pi_k) + \sum_{j=1}^{J} \log(f_{ijk}) \right\} \tag{5}$$

which corresponds to a weighted log-likelihood. Conditional on updated weights $w_{ik}^{(r)}$, we maximize $Q(\cdot)$ with respect to $\boldsymbol{\theta}$ to obtain updated ML estimates $\hat{\boldsymbol{\theta}}^{(r+1)}$. The estimated parameters are the solutions of the following M-step equations:

$$\frac{\partial Q}{\partial \pi_l} = \sum_{i=1}^{n} \left\{ \frac{w_{il}}{\pi_l} - \frac{w_{iK}}{\pi_K} \right\} = 0, \quad l = 1, \ldots, K \tag{6}$$

$$\frac{\partial Q}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \sum_{l=1}^{K} w_{il} \frac{\partial}{\partial \boldsymbol{\theta}} \left[\sum_{j=1}^{J} \log(f_{ijl}) \right]. \tag{7}$$

where θ denotes the model parameter vector. Solving the first equation we obtain $\hat{\pi}_l^{(r)} = \sum_{i=1}^{n} w_{il}^{(r)}/n$ which represents a well known result from ML in finite mixtures. Given $w_{il}^{(r)}$, (7) are weighted sums of likelihood equations for J independent GLMs. Since closed form solutions of the equations in (7) are generally unavailable, we use a standard Newton–Raphson algorithm.

3 Testing Independence Hypothesis

Testing for independence in mixed effect models is not an easy task: dependence arises through unobservable latent effects which are assumed to influence the observed outcomes through an effect included in the linear predictor. Under Gaussian assumptions on the random effect distribution, dependence is represented by linear correlation and suitable parameterizations may help specify the linear predictor as explicitly depending on the correlation coefficient between random effects. When a discrete mixing distribution is employed, null correlation does not necessarily imply independence and the correlation coefficient is not an explicit parameter in model specification. Furthermore, in the case of parametric (Gaussian) random effects the independence model is nested within the dependence model, thus standard LRT based tools may be used. When a Finite Mixture model is employed, the test for independence is not carried out under the same fixed marginal distributions obtained in case of independence. This is mainly due to the fact that the model under independence may not occur as a special case of the dependence model, since the marginal distributions under independence are estimated via several univariate models which are possibly different in the number of components, locations and/or masses from the dependence model. Let us consider two outcomes, say Y_1 and Y_2, with random effects u_1, u_2, and a number of fixed effects equal to $m_1 + 1$ and $m_2 + 1$, respectively. If K_1 and K_2 locations are used for each outcome, the total number of parameters for the independence model is equal to $d = 2(K_1+K_2-2)+(m_1+m_2+2)$, while, if K common support points are used for the bivariate model, the number of parameters is $d = 3(K-1)+(m_1+m_2+2)$. For appropriate choices of K, K_1 and K_2, the latter model could be more parsimonious and does not reduce to the former.

To understand the nature of the stochastic dependence between responses an estimate of the correlation ρ between random effects is mandatory. In the parametric mixing context numerical integration techniques (mainly ASQ) have been shown to provide consistent and reliable estimates for ρ (see Rabe-Hesket et al. 2005).

In the semiparametric context, on the other hand, ρ can be obtained as a by-product of the adopted estimation procedure. Although, considering the 'rough' nature of the estimated mixing distribution together with the reduced number of estimated locations and the extreme sensitivity to outlying observations, the corresponding estimate may be unreliable, especially in presence of non robust models (such as mixed effect logistic models for binary responses) (see Smith and Moffatt 1999, Alfó et al. 2011). In the multivariate Poisson context, the corresponding

estimates are approximately unbiased (see Alfó and Trovato 2004). In this context the LRT-statistic has a non standard distribution. To test the null (independence) hypothesis, where g, g_1 and g_2 are probability density function,

$$\begin{cases} H_0 : g(\mathbf{u}_i) = g_1(u_{i1})g_2(u_{i2}) \\ H_1 : g(\mathbf{u}_i) \neq g_1(u_{i1})g_2(u_{i2}) \end{cases} \tag{8}$$

the LRT statistic must be bootstrapped, i.e., a bootstrap sample is generated from the mixture density estimated under the null hypothesis

$$\sum_{k=1}^{K_1}\sum_{l=1}^{K_2} f_1\left(\mathbf{y}_{i1} \mid \widehat{\theta}_{i1k}\right) f_2\left(\mathbf{y}_{i2} \mid \widehat{\theta}_{i2l}\right) \widehat{\pi}_{1k}\widehat{\pi}_{2l} \tag{9}$$

and the value of $-2\log(\xi) = -2\left(\ell_{ind} - \ell_{dep}\right)$ is computed for B bootstrap samples after fitting the independence and the dependence models, both with fixed number of components, say K_1 and K_2 and K for the univariate and the bivariate models. The replicated values are then used to assess the distribution of $-2\log(\xi)$ under the null hypothesis.

4 Simulation Study

In this Section, the performance of AIC and BIC, and of the LRT are analyzed to verify whether they can be used to discriminate between the dependence and independence models when varying degrees of dependence are considered.

Random effects have been drawn from a Multivariate Normal Distribution:

$$u_i \sim \text{MVN}(\mathbf{0}, \Sigma) \qquad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \quad \rho \in \{0.00, 0.30, 0.70\}.$$

In each study, 250 samples with varying size ($n \in \{500, 1000\}$) have been used.

Two cases have been considered for the response variables. In the first one, a Poisson response with an endogenous binary selection covariate, i.e.:

$$Y_1|u_1, X_1, Y_2 \sim Poi(\lambda) \qquad Y_2|u_2, Z_1 \sim Bin(1, \pi)$$

where

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 Y_2 + u_1 \qquad \text{logit}(\pi) = \gamma_0 + \gamma_1 Z_1 + u_2$$

where X_1 and Z_1 have been drawn from a Uniform distribution on $[-1, 1]$. The (conditional) log-likelihood in this case can be written as:

$$L(\cdot \mid y_2) = \prod_{i=1}^{n} \int_{\mathscr{U}} f(y_{i1} \mid \mathbf{x}_i, y_{i2}, \mathbf{u}_i) d\mathscr{G}(\mathbf{u}_i)$$

Table 1 First simulation study: Proportion of samples where the dependence model is preferred. Percentiles of the correlation estimates

ρ	n = 500			n = 1,000		
	0.00	0.30	0.70	0.00	0.30	0.70
AIC	0.58	0.93	1.00	0.51	0.93	1.00
BIC	0.37	0.84	0.98	0.28	0.82	0.99
b-LRT	0.07	0.80	0.96	0.06	0.82	0.98
$\bar{\hat{\rho}}$	0.052	0.585	0.853	0.008	0.572	0.837
$\hat{\rho}_{0.1}$	−0.464	0.360	0.616	−0.635	0.359	0.666
$\hat{\rho}_{0.9}$	0.787	0.871	0.994	0.607	0.812	0.901

only if Y_2 and u_1 are assumed to be independent. If this is not the case endogeneity bias may arise and the *primary* model must be supplemented by a *secondary* model accounting for regressor endogeneity.

The number of components for each mixture varied from 2 to 20.

The LRT has been performed through $B = 100$ bootstrap samples drawn under the null hypothesis of independence between outcomes; the number of components, say K_1, K_2 and K, representing the null and the alternative hypotheses (independence, dependence) have been chosen in each sample via BIC, while model parameters under the null and the alternative have been estimated in the resamples by fixing the corresponding number of components to K_1, K_2 and K.

In Table 1 the proportion of samples where the dependence model is preferred over the independence model according to AIC, BIC and the bootstrapped LRT of size 0.05 have been reported together with the average of the estimates of the correlation coefficients derived from the dependence model and the 10th and 90th percentiles of its distribution over the 250 samples. In each sample/resamples, we employed 20 different (randomly chosen) starting values.

It is worth noting that both AIC and BIC tend to prefer the dependence model, even when the independence model represents the *true* data generating process. This could be ascribed to the fact that both are quite reliable in selecting the *right* number of components (if any) in finite mixtures, but tend to favour parsimonious models, thus preferring the dependence model. On the other side, the LRT is more powerful in choosing between the two models, even with small sample sizes ($n = 500$).

A behaviour consistent with other findings in the literature (see Smith and Moffatt 1999, Alfó et al. 2011) can be observed for the correlation coefficient estimates, with a large variability of the estimates, a high percentile range, and a clear tendency to overestimation when the *true* ρ approaches the corresponding bounds.

In the second case study, we considered Poisson and Exponential responses

$$Y_1|u_1, X_1 \sim Poi(\lambda) \qquad Y_2|u_2, Z_1 \sim Exp(\theta)$$

where

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + u_1 \qquad \log(\theta) = \gamma_0 + \gamma_1 Z_1 + u_2$$

Table 2 Second simulation study: Proportion of samples where the dependence model is preferred. Percentiles of the correlation estimates

ρ	n = 500			n = 1,000		
	0.00	0.30	0.70	0.00	0.30	0.70
AIC	0.95	1.00	1.00	0.94	0.99	1.00
BIC	1.00	1.00	1.00	0.98	1.00	1.00
b-LRT	0.065	0.83	0.98	0.055	0.86	0.99
$\bar{\hat{\rho}}$	0.005	0.296	0.653	0.002	0.305	0.656
$\hat{\rho}_{0.1}$	−0.146	0.055	0.437	−0.079	0.202	0.626
$\hat{\rho}_{0.9}$	0.161	0.433	0.869	0.129	0.391	0.804

This could be the case, for instance, of health care utilization and expenditures data jointly determined by, say, health care status or propensity to use a given service (maybe on an out-of-pocket basis). In Table 2 the results of the simulation study for the second case have been displayed. The findings which can be derived are coherent with those obtained for the first case study, with some differences. Here the AIC and BIC behaviours are worse than the performance in the first case study; this could be due to the reduced number of locations that a finite mixture of logistic models usually needs for the independence model; thus, in this case, with a continuous (exponential) outcome the performance of the penalized likelihood criteria is poor and likely a high number of locations are needed to fit the univariate marginal distributions. In this case, the dependence model is always preferred over the independence one. This is not true when a bootstrapped LRT is employed: the LRT performs fairly well, and is quite powerful, even if the number of resamples under the null hypothesis is rather limited (i.e., $B = 100$).

As long as the correlation estimates are considered, contrary to what experienced in the first simulation study, $\hat{\rho}$ values show limited variability as well as satisfactory mean estimates, which are quite close to the corresponding *true* values, regardless of whether they are near to the bound (i.e., $\rho = 0.7$).

5 Conclusions

Random effect models can be easily adapted to handle multivariate mixed data. While AIC or BIC can be used to choose the number of components in a finite mixture approach, they are of little help to *test* for independence.

The LRT performs better than penalized likelihood criteria in testing for dependence; in fact, the bootstrapped LRT statistic has been shown to perform quite well in all analyzed situations, even with small sample sizes ($n = 500$) and small number of bootstrap resamples ($B = 100$). Nonetheless AIC and BIC show a quite reliable behaviour, at least on average, when the estimation of regression coefficients and random effect correlation are considered. The estimates of ρ are quite good in the case of Poisson–Exponential outcomes, with moderate variability. The results for

the Poisson–Bernoulli case are not as good, showing a very high variability in the simulation study. This suggests the use of some caution when dealing with Bernoulli responses.

References

Alfó, M., Maruotti, A., & Trovato, G. (2011). A finite mixture model for multivariate counts under endogenous selectivity. *Statistics and Computing, 21*, 185–202.

Alfó, M., & Trovato, G. (2004). Semiparametric mixture models for multivariate count data, with application. *Econometrics Journal*, *7*, 1–29.

Chib, S., & Winkelmann, R. (2001). Markov chain monte carlo analysis of correlated count data. *Journal of Business and Economic Statistics*, *19*, 428–435.

Crouch, E. A. C., & Spiegelman, D. (1990). The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(t)\exp(-t^2)dt$: Application to logistic-normal models. *Journal of the American Statistical Association*, *85*, 464–469.

Gueorguieva, R. (2001). A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Statistical Modelling*, *1*, 177–193.

Harry, J. (1997). *Multivariate models and multivariate dependence concepts*. London: Chapman & Hall/CRC.

Heckman, J. J., & Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models of duration. *Econometrica*, *52*, 271–320.

Karlis, D., & Meligkotsidou, L. (2006). Multivariate poisson regression with covariance structure. *Statistics and Computing*, *15*, 255–265.

Kiefer, J., & Wolfovitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, *27*, 887–906.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, *73*, 805–811.

Munkin, M. K., & Trivedi, P. K. (1999). Simulated maximum likelihood estimation of multivariate mixed-poisson regression models, with application. *Econometrics Journal*, *2*, 29–48.

Rabe-Hesket, S., & Skrondal, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata Journal*, *2*, 1–21.

Rabe-Hesket, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, *128*, 301–323.

Smith, M. D., & Moffatt, P. G. (1999). Fisher's information on the correlation coefficient in bivariate logistic models. *Australian and New Zealand Journal of Statistics*, *41*, 315–330.

Size and Power of Tests for Regression Outliers in the Forward Search

Francesca Torti and Domenico Perrotta

Abstract The Forward Search is a method for detecting masked outliers and for determining their effect on models fitted to the data. We have estimated the actual statistical size and power of the Forward Search in regression through a large number of simulations, for a wide set of sample sizes and several dimensions. Special attention is given here to the statistical size. The work confirms for regression the excellent Forward Search properties shown in the multivariate context by Riani et al. (Journal of the Royal Statistical Society. Series B 71:1–21, 2009).

1 Introduction

The Forward Search (FS) is a general method for detecting unidentified subsets and masked outliers and for determining their effect on models fitted to the data. The key concepts at the basis of the FS originated from the work of Hadi (1992), but the power of the method was increased by Atkinson and Riani (2000) and Atkinson et al. (2004) through the idea of diagnostic monitoring, which was applied to a wide range of regression and multivariate statistical techniques. Unlike most robust methods (e.g., Rousseeuw and Leroy 1987; Maronna et al. 2006) in the FS the amount of trimming is not fixed in advance, but is chosen conditionally on the data. Many subsets of the data are fitted in sequence. As the subset size increases, the method of fitting moves from very robust to highly efficient likelihood methods.

The FS has been applied in many contexts. In particular Riani et al. (2008) used the FS to analyse bivariate trade data arising in the market between the European Union and the rest of the world. In such contexts, outliers are the key objective of the analysis because some of them may correspond to fraudulent transactions. Unlike other situations where the focus is on the power of the statistical method, here it is crucial to control the size of the outlier tests, to avoid submitting to the anti-fraud services an intractable number of (often irrelevant) cases to inspect.

The actual size and power of the FS in the multivariate context, where the tested statistic is the Mahalanobis distance, were studied by Riani et al. (2009). However no equivalent evaluation has been done for regression. This work fills this gap with

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_42,

an extensive evaluation of the actual size and power of the FS outlier test in the regression context. We have compared the nominal size α and power with the actual ones produced by the FS, following the benchmark scheme of Riani et al. (2009) in terms of number of simulations, subset sizes, number of variables, contamination levels, power measures, etc., The results obtained show that the empirical α is very close to the nominal one. We have also assessed the size and the power of the best existing outlier techniques based on the Least Trimmed of Squares (LTS) and Least Median of Squares (LMS) (Rousseeuw 1984) estimation methods. These results have been compared with those obtained for the FS and a traditional backward strategy based on repeated removal of the observation with significant deletion residual and highest Cook distance, along the lines of Marasinghe (1985). Although it is well known that in general backward approaches are subject to masking and swamping problems, in practice the method produced satisfactory results on bivariate regression trade data, analysed with that method in the Joint Research Centre of the European Commission (EC, JRC). This is why we addressed more formally its statistical performance.

The FS methodology for regression is briefly introduced in Sect. 2. In Sect. 3 we describe the benchmark experiment and the results on the statistical size. We introduce the related work on power in Sect. 4. The benchmark was also used to monitor the time complexity of the methods, to evaluate applicability to real world problems (Sect. 5). The overall results are briefly discussed in the final section.

2 The Forward Search in Regression

The FS for regression is given book-length treatment by Atkinson and Riani (2000). The basic idea is to start from a small robustly chosen subset of data and to fit subsets of increasing size, in such a way that outliers and subsets of data not following the general structure are revealed by diagnostic monitoring. In the regression context we have one univariate response Y and p explanatory variables $X_1, \ldots, X_p$ satisfying

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \tag{1}$$

under the usual assumptions on the linear model and the errors ϵ_i in particular. Let $\hat{\beta}(m)$ be the estimate of the $(p+1)$-dimensional parameter vector $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ obtained by fitting the regression hyperplane to subset $S(m)$. From this estimate we compute n squared regression residuals

$$e_i^2(m) = [y_i - \{\hat{\beta}_0(m) + \hat{\beta}_1(m)x_{i1} + \cdots + \hat{\beta}_p(m)x_{ip}\}]^2 \qquad i = 1, \ldots, n \tag{2}$$

which are used for defining the new subset $S(m+1)$. The search starts from an outlier-free subset of $p+1$ observations satisfying the LMS. To detect outliers we now examine the minimum deletion residual amongst observations not in the subset

$$r_{\min}(m) = \min \frac{|e_i(m)|}{s(m)\sqrt{[1 + x_i^T \{X^T(m)X(m)\}^{-1} x_i]}} \quad \text{for } i \notin S(m), \qquad (3)$$

where $s(m)$ is the square root of the unbiased estimate of the residual variance $\sigma^2 = E\{y_i - E(y_i)\}^2$ computed from the observations in $S(m)$, $x_i = (x_{i1}, \ldots, x_{ip})^T$ is the ith row of the design matrix X and $X(m)$ is the block of X with rows indexed by the units in $S(m)$. Inferences about the existence of outliers require envelopes of the distribution of $r_{\min}(m)$ (Atkinson and Riani 2006). In absence of outliers (3) will progress stably within the envelopes otherwise, when the observations outside the subset are outliers, it will show a jump exceeding the upper envelopes. The exact rule to find significant exceedances follows the same scheme of Riani et al. (2009) for the multivariate context.

3 Benchmark Setup and Results on the Statistical Size

The benchmark setup was the same for the four linear regression methods assessed (FS, LMS, LTS and backward iterative method). It was run using Matlab and its Statistical toolbox. LTS and LMS were run using code taken from the Matlab Library for Robust Analysis (LIBRA) developed by Mia Hubert (http://wis.kuleuven.be/stat/robust/LIBRA.html). For the FS we used the Forward Search for Data Analysis (FSDA) toolbox of Marco Riani et al. (http://www.riani.it/MATLAB.htm). The backward method was ported to MATLAB from the original SAS implementation developed in the JRC.

The FS was run at its standard nominal significance level $\alpha = 0.01$. The subset from which the FS starts progressing was found by running LTS on up to 10,000 sub-samples. We assessed a variant of LTS and LMS, available in LIBRA, which adds a final re-weighting step to the standard algorithms to increase their efficiency (Rousseeuw and Leroy 1987). To obtain an overall size $\alpha = 0.01$, each LTS/LMS session was run at the Bonferroni-corrected size $(1 - 0.01/n)$, where n is the sample size. Moreover, the initial robust LTS/LMS estimator was found by extracting 10,000 sub-samples. This choice is motivated by the fact that the results obtained with 10,000 sub-samples were considerably different from those obtained with only 1,000 sub-samples, but very close to those obtained with 50,000 sub-samples. The backward method was also run to achieve the overall $\alpha = 0.01$ with Bonferroni-corrected individual tests.

Independently from the method, each benchmark experiment was based on 10,000 replications, i.e., on 10,000 sets of data. To check if this number of replications was sufficient to achieve a reasonable accuracy on the size estimates, we have also performed few benchmark experiments with 50,000 replications. The results were very close, as Table 1 shows for the FS case, and therefore we have limited the benchmark experiment to 10,000 sets of data. Simulations were made for five sample size values $n = 1000, 500, 400, 200, 100$ and for four values for the number

Table 1 Empirical size of the nominal 1% outlier test for the forward search based on 50,000 (left) and 10,000 (right) sets of data, for different sample sizes n and number of explanatory variables p

n	$p = 2$	$p = 5$	$p = 10$	$p = 1$	$p = 2$	$p = 5$	$p = 10$
100	0.0105	0.0114	0.0262	0.0118	0.0099	0.0127	0.0283
200	0.0115	0.0106	0.0173	0.0171	0.0108	0.0112	0.0173
400	0.0122	0.0105	0.0154	0.0178	0.0121	0.0109	0.0151
500	0.0126	0.0112	0.0149	0.0175	0.0132	0.0095	0.0126
1,000	0.0130	0.0115	0.0162	0.0167	0.0138	0.0117	0.0147

of explanatory variables $p = 1, 2, 5, 10$. Results were declared significant if *at least* one outlier was detected.

The null hypothesis to test is that data are normal, $H_0 : \epsilon_i \sim N(0, \sigma^2)$, therefore the response variable has been generated from a normal distribution in each replication of the benchmark experiment. The values of the explanatory variable were fixed for all the 10,000 replications (sets of data). Such values were also generated from a normal distribution. Note that different choices would lead to different leverage problems and therefore to different benchmarks.

To start with the results, the Table 1 (left) reports the FS empirical α obtained with 50,000 replications, i.e., sets of simulated data. The corresponding results on 10,000 replications are on the right table. The difference between the comparable values in the two tables is negligible. On the contrary, a similar check on 1,000 replications gave results considerably different and probably inaccurate. Therefore, we ran all other benchmark experiments to 10,000 sets of data.

Figure 1 reports for $p = 1$ (top left plot), $p = 2$ (top right plot), $p = 5$ (bottom left plot) and $p = 10$ (bottom right plot) the actual α of FS (solid line), LTS (dotted line) and LMS (dashed line). LMS has an empirical α which is constantly lower than FS and LTS. In particular it is surprising that the LMS empirical α is even lower than the nominal size of 1% for reasons that must be investigated. Another aspect common to all p values is that the LTS α tends to decrease with n, and for small sample sizes it assumes rather high values. Finally, in the four plots the FS empirical α is just above the nominal value: in particular for $p = 2$ and $p = 5$ the maximum value is 0.013, while for $p = 1$ and $p = 10$ the maximum values are 0.0178 and 0.0283 respectively.

We dedicated particular care to the assessment for small samples. One reason is that the FS outlier tests are based on theoretical envelopes that are known to have many good properties (e.g., they account for the multiplicity of the tests), but till now it has never been shown how good they are in practice for small sample sizes. A second reason is that in international trade analysis there are problems requiring few years of monthly trade aggregates, i.e., datasets formed by less than 50 entries each. For these reasons we assessed the test size for datasets of sample sizes multiple of 5 in the range $15 \leq n < 100$. We limited this specific benchmark to one dependent variable ($p = 1$). This choice is because the deletion residual tests are based on $(n - p - 1)$ degrees of freedom and it is not reasonable to reduce to too few degrees

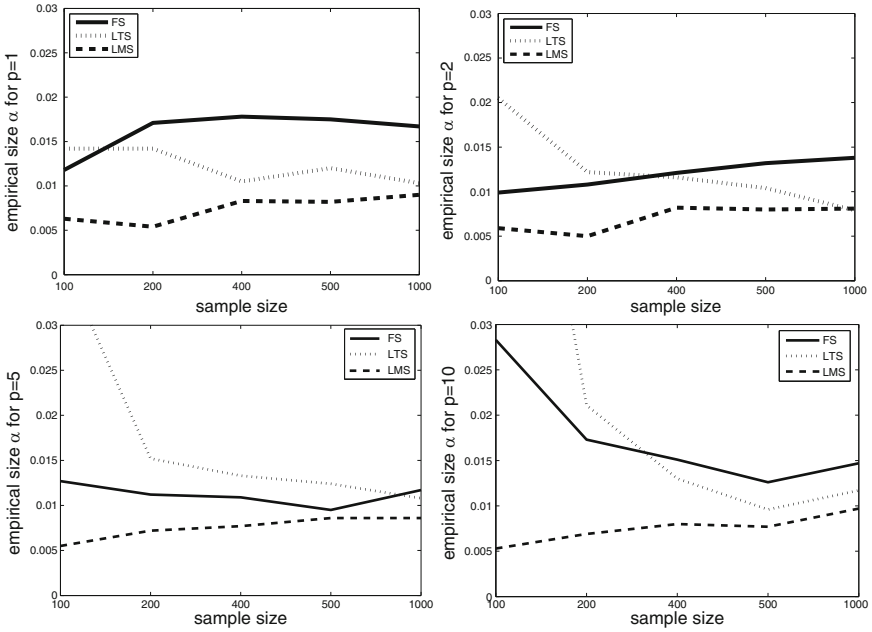


Fig. 1 Empirical size of forward search (*solid line*), LTS (*dotted line*) and LMS (*dashed line*) for a nominal 1% outlier test for $p = 1$ (*top left plot*), $p = 2$ (*top right plot*), $p = 5$ (*bottom left plot*) and $p = 10$ (*bottom right plot*) and for sample sizes between 100 and 1,000

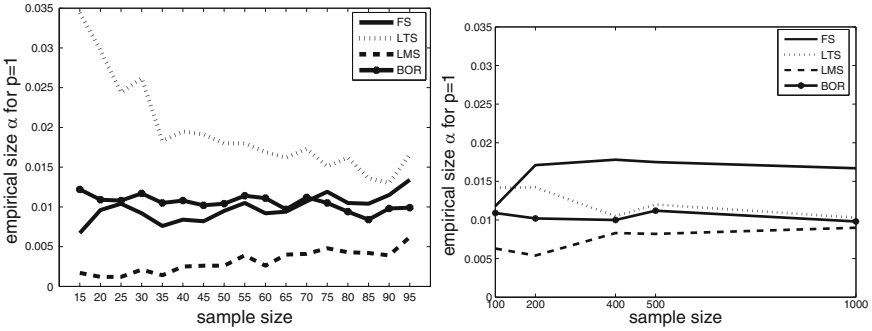


Fig. 2 Empirical size α of forward search (*solid line*), LTS (*dotted line*), LMS (*dashed line*) and the backward method (BOR, *line with stars*) when $p = 1$ and n among 15 and 95 (*left plot*) and between 100 and 1,000 (*right plot*)

of freedom. Besides in trade analysis typically the trade value is regressed only over the quantity, and this is the case of main interest for us currently.

The left plot of Fig. 2 reports the α estimated for the FS, LTS, LMS and the backward technique for small samples. In order to compare the performances for small samples with those for bigger sizes, on the right we present again the top left plot of Fig. 1 with the results of the backward method superimposed. For small n

(left plot) the FS (solid line) shows an α increasing from 0.6% to 1.13%, while, for larger n (right plot), it increases and stabilizes around 1.6%. The LTS (dotted line) for n small has a simulated α with high variability which assumes values greater than 3%, while, for larger n, it stabilizes around a value just above 1%. The LMS (dashed line) always has very low α values, but for small n the size is even less than 0.5%. The backward method (line with stars) is stable around 1% for all values of n. This is a surprisingly good result for that method.

4 A Few Anticipatory Remarks on Statistical Power

The study of power has produced many more results that will be discussed at length in a separate publication. However we anticipate here how the power benchmark was conceived in relation to the size and give a claim on the overall achievement.

To compare the power performance of the various outlier tests we need them to have at least approximately the same size. Based on the plots in Fig. 1, size curves are closer for the combination ($p = 5, n = 500$). The power has been mainly assessed for that combination. Here the LTS has the highest empirical α (0.0124), followed by the FS (0.0095) and finally by the LMS (0.0086). For small samples (right plot of Fig. 2) we can see that the simulated α of LTS is constantly higher than the one of the FS, which is very close to that of the backward method and higher than the one of the LMS. However there is no particular value of n for which the curves are particularly close and we have decided to monitor the power for $n = 50$. Here the LTS has the highest α (0.0180), followed by the backward method (0.00197), by the FS (0.0095) and finally by the LMS (0.0026).

We organised the benchmark on power with the same settings as those for size, with a contamination scheme for the response variable which is rather standard in the robust statistics literature: 5% of the response values were subject to a shift increasing were shift of a size increasing from 1 to 7. Five different measures of power have been computed and analysed: (i) average power, i.e., the average, over all iterations, of the number of true detected outliers with respect to the contaminated observations, (ii) simultaneous power, i.e., the average, over all iterations, of the number of detected (both true and false) outliers, (iii) family wise error rate, i.e., the average number of iterations where at least one false outlier has been detected, (iv) false discovery rate, i.e., the average, over all iterations, of the number of false detected outliers with respect to all the detected (both true and false) outliers, (v) proportion of declared outliers in good data, i.e., ratio between the average number of false detected outliers with respect to the number of non-contaminated observations.

To anticipate these power results we can say that for $n > 100$ the FS has clearly shown superior performance for any sample dimension, followed by LTS and LMS. For small samples the FS power is still excellent but the results are not so clearly interpretable in relation to the other methods, because of the less neat size results discussed above. Material substantiating these claims is available from the authors.

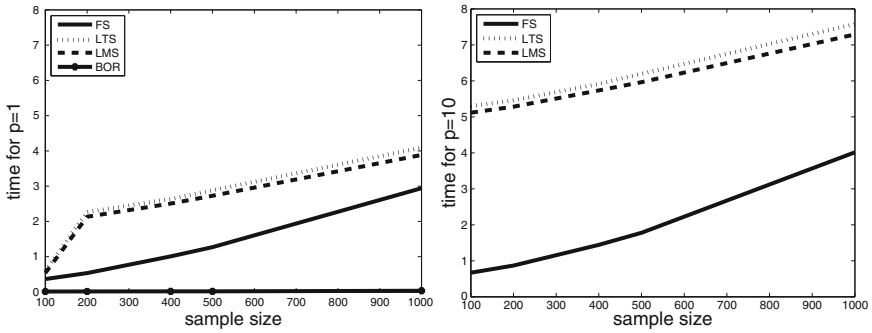


Fig. 3 Estimated elapsed time in seconds for forward search (*solid line*), LTS (*dotted line*), LMS (*dashed line*) when the sample size n is among 100 and 1,000 and $p = 1$ (*left plot*) and $p = 10$ (*right plot*). For $p = 1$ the results of the backward method (BOR) are also represented with a dotted line with stars. The LMS/LTS runs were based on Mia Hubert's Matlab implementation (LIBRA)

5 Considerations on Computational Efficiency

We have also monitored the CPU time in seconds needed to execute the four methods, for a range of sample sizes n and number of explanatory variables p. For each technique, the estimated computational time for given n and p is the mean of the CPU time values monitored for each of the 10,000 replications. The results are summarised in Fig. 3. The backward approach (line with stars in the left plot) is based on a computationally simple and efficient algorithm, which is therefore very fast compared to the other methods. Among the robust approaches, the FS is the fastest: the average time curve lies below LTS and LMS for all sample size values n and even for $p = 10$. Although the FS finds an initial robust estimator using the LTS algorithm, the LTS is applied only once to find a small and sub-optimal subset of $p + 1$ observations (at most 10,000 random subsets are considered). This explains why the call to LTS does not appreciably affect the FS computational time.

The LTS time estimates are those for the standard re-weighted LTS. We have also made a statistical and time assessments of a fast version of LTS (FAST-LTS algorithm), that is known in the literature for its better computational performances. Its size and power were similar to the standard LTS. Time-wise, the performances improved with respect to LTS but they were still not comparable with the FS.

6 Discussion

The results show that the Forward Search tests have good statistical size for a wide range of sample sizes ($15 \leq n \leq 1000$) and dimensions ($1 \leq p \leq 10$). The good performance for small sample sizes confirms the excellent properties of the

theoretical envelopes introduced by Riani et al. (2009). The results also show a very good size of LMS for the full range of dimensions n, and a good size for LTS but for $n > 100$. The interpretation of the corresponding results for smaller n is not so neat: for $15 \leq n \leq 100$ the statistical size is the smallest for LMS, followed by the Forward Search and LTS.

These results on the FS size will be complemented in a separate publication by the corresponding power results, that we anticipate to be excellent.

An empirical analysis of the computational complexity of the four methods has finally shown that the Forward Search in practice is faster (excluding the backward method, that is not comparable to the other methods in terms of statistical performances). This is because LMS/LTS have a severe combinatorial problem to overcome, while the Forward Search can use such methods for the choice of the initial subset with a reduced number of random subsets.

Acknowledgements The work was partially supported by the Joint Research Centre of the European Commission, under the institutional work-programme 2007–2013 of the research action "Statistics and Information Technology for Anti-Fraud and Security".

References

Atkinson, A. C. & Riani, M. (2000). *Robust diagnostic regression analysis*. New York: Springer.

Atkinson, A. C. & Riani, M. (2006). Distribution theory and simulations for tests of outliers in regression. *Journal of Computational and Graphical Statistics*, *15*, 460–476.

Atkinson, A. C., Riani, M., & Cerioli, A. (2004). *Exploring multivariate data with the forward search*. New York: Springer.

Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society Series B*, *54*, 761–771.

Marasinghe, M. G. (1985). A multistage procedure for detecting several outliers in linear regression. *Technometrics*, *27*, 395–399.

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics: Theory and methods*. Chichester: Wiley.

Riani, M., Atkinson, A. C., & Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society. Series B*, *71*, 1–21.

Riani, M., Cerioli, A., Atkinson, A. C., Perrotta, D., & Torti, F. (2008). *Fitting robust mixtures of regression lines to European trade data, in mining massive datasets for security applications*. IOS Press, Amsterdam.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association 79*, 871–880.

Rousseeuw, P. J. & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.

Using the Bootstrap in the Analysis of Fractionated Screening Designs

Anthony Cossari

Abstract In recent years the bootstrap has been favored for the analysis of replicated designs, both full and fractional factorials. Unfortunately, its application to fractionated designs has some limitations if interactions are considered. In this paper the bootstrap is used in a more effective way in the analysis of fractional designs, arguing that its application is closely connected with the projective properties of the design used. A two-stage approach is proposed for two-level orthogonal designs of projectivity $P = 3$, both replicated and unreplicated, which are especially useful in screening experiments. This approach combines a preliminary search for active factors with a thorough study of factor effects, including interactions, via a bootstrap analysis. It is especially useful in non-standard situations such as with nonnormal data, outliers, and heteroscedasticity, but depends heavily on the assumption of factor sparsity. Three examples are used for illustration.

1 Introduction

Bootstrapping is a popular data-resampling technique for making analyses in a wide range of problems. Basically, the bootstrap generates (with replacement) R new samples from the observed one in order to build, for any statistic of interest, the empirical bootstrap distribution that will be conveniently summarized to make inferences, for example to obtain standard errors or confidence intervals. A comprehensive account of bootstrap methods can be found, e.g., in Davison and Hinkley (1997). Recently, Kenett et al. (2006) favored the application of the bootstrap for the analysis of both full factorial and fractionated designs as an alternative to a standard regression approach in cases where regression may fail, such as in the presence of nonnormal data, outliers, heteroscedasticity, or model misspecification. Moreover, they introduced a valuable diagnostic tool to reveal such special cases, based on the comparison between the regression and the bootstrap standard errors. Their proposal have bridged a surprising gap in the design of experiments literature which has rarely been concerned with bootstrap-based analysis methods. However, the applicability of the bootstrap is restricted to designs, either full or

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_43,

fractional factorials, with replicated data. This ensures that the data are resampled in a multi-sample fashion, providing the most effective way of bootstrapping.

In the analysis of fractionated designs, model misspecification may occur due to aliasing of the possible effects. For example, a rather common approach to deal with the aliasing patterns is to just neglect all interactions among factors, thus considering a model with the main effects only. Such approach, however, may lead to wrong conclusions if some interactions do matter. Instead, potentially important interactions should be included in the model to make a valid and reliable analysis. Kenett et al. (2006) provide an effective way to test for model misspecification via their diagnostic tool, in order to suggest the most appropriate model to use. However, their procedure has some drawbacks when interactions are to be included in the model. Going beyond the method by Kenett et al. (2006) for revealing model inadequacies, in this paper a bootstrap-based analysis of fractionated designs is considered, which is closely connected with the property of design projectivity defined in Box and Tyssedal (1996). A two-stage approach is proposed for two-level orthogonal designs of projectivity $P = 3$, both unreplicated and replicated. These types of designs are particularly important in screening experiments under the usual condition of factor sparsity (e.g., see Box and Meyer, 1993). This approach allows to find significant interactions as well as main effects, but relies heavily on the factor sparsity assumption. Moreover, it is especially useful with data from nonstandard situations. The procedure combines an initial search for active factors with a bootstrap analysis of a full factorial. The paper is organized as follows. In Sect. 2 a real example that has motivated this work is discussed. In Sect. 3 the proposed two-stage approach for the analysis is outlined, while Sect. 4 shows its application with three examples. Finally, Sect. 5 gives a conclusion.

2 Motivating Example

Kenett et al. (2006) use an experiment on wave soldering to illustrate the application of the bootstrap in testing model misspecification for replicated fractional factorials. The experiment is based on a familiar 2^{7-3} design with three replicates, used for studying the influence of seven factors, denoted by A–G, on the number of defects in the soldering process. Standard regression analysis, which is based on the normality assumption, is not appropriate for such count data, but it may be applied to the square root of the data. Although simple and attractive, this procedure has some limitations, as explained by Wu and Hamada (2000, p. 563), who have used the wave soldering experiment to illustrate the application of the generalized linear model approach for the analysis of this type of data. The design is of resolution *IV*, which implies that the main effects are confounded with three-factor interactions. A model including only the main effects is commonly employed with this type of design, based on the assumption that three-factor interactions are negligible. Kenett et al. (2006) use their diagnostic procedure to test for such a main-effect model in the soldering experiment. Moreover, confidence intervals for

each parameter in the model are calculated for analysis purposes. The main-effect model turns out to be not adequate. Based on this result, Kenett et al. (2006) suggest considering a model containing the main effects emerged as significant in the main-effect analysis, namely those of factors A, B, C and G, together with all interactions among these factors. The diagnostic test supports the validity of this model, which is therefore recommended for the analysis. However, a possible disadvantage of a main-effect model, such as that initially tested in the soldering experiment, is that it may lead to wrong analyses if the assumption that interactions are negligible is not true, even though the interactions neglected are of the third order (e.g., see Box and Meyer, 1993; Jacroux, 2007). For instance, main effects which are not real may be selected as being significant, or factors involved in significant interactions may be completely missed. As a consequence, a more complete model with both main effects and interactions should not be derived from the results of a preliminary main-effect analysis. Instead, it should be fitted with an analysis method that can incorporate directly such a model.

Hence, while the test procedure introduced by Kenett et al. (2006) is perfectly suitable for judging the appropriateness of a main-effect model, it seems to have some limitations when a model with interactions has to be taken into account, resulting in questionable analyses. These limitations has motivated the need to riconsider the general problem of making effective use of bootstrapping in the analysis of fractionated designs, when such analysis aims at studying interactions as well as main effects. An approach to analysis has been derived for designs of projectivity $P = 3$, and is outlined in the next section.

3 Approach to Analysis

The approach presented in this section takes advantage of the property of projectivity for two-level orthogonal designs. The projective properties of fractionated designs have been investigated by several authors, including Box and Hunter (1961) who originated the concept, and later, e.g., Lin and Draper (1992), Cheng (1995), and Box and Tyssedal (1996, 2001).

According to the definition of projectivity introduced by Box and Tyssedal (1996), a two-level orthogonal design with N runs and k factors is of projectivity P if every subset of P factors out of k gives a 2^P full factorial design, possibly with some or all of its points replicated. The projectivity of a design is especially important when the analysis is made under the assumption of factor sparsity. Factor sparsity implies that, out of the k factors investigated, only a few of them will be the active ones, usually a small fraction. A factor, say A, is active if either the main effect of A or an interaction involving A or both are significant. Such assumption typically holds when fractionated designs are used for screening purposes at the initial stage of an investigation. An efficient approach to analysis of screening designs is to first find the most likely active factors, and then study thoroughly their potential effects, including interactions, with further analyses or after appropriate additional

experiments. For a design of projectivity P, if up to P factors are indicated as likely active via some preliminary analysis, the projected full factorial design onto these active factors gives the opportunity to fit the most complete model, containing all the possible interactions in addition to the main effects, to estimate all these effects in greater detail. Most screening designs of practical use under factor sparsity have projectivity $P = 2$, some have projectivity $P = 3$. Therefore it is important to choose a design of the highest projectivity $P = 3$ to gain from its estimation possibilities.

In the familiar 2^{k-p} series, the designs of projectivity $P = 3$ are all those that have resolution IV (e.g., see Box and Hunter, 1961). Among such designs, the most relevant ones in a screening context are the 16-run 2^{6-2}, 2^{7-3}, and 2^{8-4} designs, for studying 6, 7, and 8 factors respectively. Notice that a 2^{7-3} plan was employed in the soldering experiment reported earlier. An extremely useful screening design of projectivity $P = 3$ is the popular 12-run Plackett-Burman design (see Plackett and Burman, 1946), which can accommodate up to 11 factors. Despite its fame of a main-effect plan due to the complexity of the aliasing pattern, its great potential to find important interactions is now widely recognized since the groundbreaking paper by Hamada and Wu (1992). Other orthogonal arrays of projectivity $P = 3$, that can be useful for screening though not very common in practice yet, are four 16-run designs introduced by Box and Tyssedal (2001). All these designs of projectivity $P = 3$ have projections in $P = 3$ or fewer factors with some or all points replicated.

Besides being useful for estimation purposes, a design of projectivity $P = 3$ may be exploited to search for the few active factors by taking full advantage of the replicated data in each projection, as first demonstrated by Tyssedal and Samset (1997), and later by Tyssedal et al. (2006) and Tyssedal (2007). In fact, each candidate subset with $P = 3$ or fewer factors projects onto a full factorial design in which replicates may be used to assess the fit to data for the given subset of factors, whichever regression model is assumed for these factors, since replicated data have the same expected value. A convenient measure of fitness is the usual pooled estimate of the error standard deviation, say $\hat{\sigma}$. If the true active factors are up to $P = 3$, their subset usually stands out with the smallest value of $\hat{\sigma}$. If the active factors are more than $P = 3$, the procedure can't generally identify them. In this case, there is usually no clear winner in the list of subsets ranked after $\hat{\sigma}$. This search method for active factors will be referred to as the projectivity-based search method. A more general, and also more elaborate, method for finding the active factors is the Box-Meyer method (Box and Meyer, 1993), which can be applied to all types of fractionated designs. Such method allows to calculate, for each factor, its posterior probability of being active, and hence declares as active those factors whose posterior probabilities stand out from the rest.

With the primary intention of using the bootstrap in the most effective way with fractionated designs, a two-stage approach to analysis is introduced for two-level orthogonal designs of projectivity $P = 3$, both replicated and unreplicated. At the first stage, a preliminary search for active factors is carried out. The projectivity-based method may be used for its simplicity and close connection with the projective properties of the design. When such method fails or gives unclear conclusions,

a supplemental screening analysis with Box-Meyer method can usually be effective in completing this stage. If the active factors selected at the first stage are up to $P = 3$, the second stage may generally be run, where the single projected design in the active factors, a replicated full factorial in three factors at most, is analyzed with the bootstrap to obtain confidence intervals for all the terms in a full model, thus making inferences for main effects and all the possible interactions. If the first stage points to more than $P = 3$ active factors, the procedure is inconclusive in that the corresponding projected design is not adequate to investigate thoroughly the effects of these active factors. However, such projected design usually gives clear indications of how it should be augmented with follow-up runs to obtain a full factorial. An adaptation of this procedure is needed for the 12-run Plackett-Burman design, when such design is unreplicated and the active factors are exactly $P = 3$. In fact, the projection onto the three active factors, whichever they are, is a 2^3 design with only four of its points replicated twice, resulting in too few distinct samples for implementing the bootstrap algorithm. In this case, four additional experiments in the unreplicated runs of the projected design will produce a 2^3 full factorial, with two replicates, suitable for bootstrapping.

It is clear that this two-stage approach depends heavily on the factor sparsity assumption. If such assumption holds, however, the procedure can effectively reveal significant factor effects including interactions via the bootstrap analysis. It can work on all types of data, thus it is especially useful in nonstandard cases, such as data which are nonnormal or have outlying observations, and heteroscedastic data.

4 Examples

Three examples are presented to illustrate the value of the proposed two-stage approach together with its limitations. All the computations were carried out with the R language. The 'boot' package by Canty and Ripley was used to implement the bootstrap algorithm.

In the first example, analysis of the count data in the soldering experiment is reconsidered. The associated 2^{7-3} design of projectivity $P = 3$, replicated three times, reduces to a 2^1 design with $3 \times 8 = 24$ replicates for each factor, a 2^2 design with $3 \times 4 = 12$ replicates for each couple of factors, and a 2^3 design with $3 \times 2 = 6$ replicates for each subset of three factors, with a total of 63 projected designs to be examined at the first stage of the procedure through the projectivity-based search method. Under the assumption of factor sparsity, we expect to find two or three active factors. Table 1 reports a representative portion of the list of subsets of factors ranked after $\hat{\sigma}$, namely the top five subsets for one factor, two factors and three factors respectively. These results show that none of the subsets is a clear winner, thereby suggesting that more than $P = 3$ factors are needed to explain the data. An additional analysis through Box-Meyer method points to the four factors A, C, E and G as likely active, with posterior probabilities which are substantially higher than those of the other factors.

Table 1 Top five subsets of lowest $\hat{\sigma}$, soldering example

One factor	$\hat{\sigma}$	Two factors	$\hat{\sigma}$	Three factors	$\hat{\sigma}$
C	33.29	A, C	30.16	A, C, E	26.75
G	36.31	C, G	30.31	A, C, G	26.75
A	36.63	B, C	31.82	A, E, G	26.75
B	36.99	C, E	33.75	C, E, G	26.75
E	37.86	C, D C, F	34.01	A, B, C	27.53

Since the active factors are more than $P = 3$, the procedure fails to study all their possible effects via the bootstrap analysis. Nevertheless, the active factors are identified correctly. By comparing the results of the two-stage approach with those of the main-effect analysis of Kenett et al. (2006), it can be argued that factor E was declared active at the first stage because it is likely involved in significant interactions, but its main effect is not significant. Analysis of Kenett et al. (2006) misses factor E, while picking the unimportant factor B with a supposed significant main effect. Indeed, any main-effect analysis on these data does fail to reveal the importance of factor E, such as the generalized linear model analysis in Wu and Hamada (2000). Clearly, this challenging experiment does not follow exactly the principle of factor sparsity, and shows the limitations of the two-stage approach when the active factors are more than $P = 3$. The projection onto the active factors A, C, E and G happens to be a 2^{4-1} half fraction with $3 \times 2 = 6$ replicates. Just one replicate of the complementary 8-run half fraction would lead to an augmented 2^4 full factorial, with eight replicated points, perfectly adequate for bootstrapping.

The second example considers a screening experiment based on a 12-run Plackett-Burman design with ten replicates, which was used in a reliability improvement study to assess the influence of 11 factors, denoted by $A - K$, on the right-censored failure times of an industrial thermostat. Besides being non-normal with right–censoring, these data also show significant heteroscedasticity. Wu and Hamada (2000, p. 531) have analyzed the thermostat experiment using a likelihood-ratio test approach with a lognormal model, and a Bayesian variable selection approach. Kenett et al. (2006) have used these data to demonstrate the value of their bootstrap-based diagnostic tool in detecting heteroscedasticity. Table 2, analogous to Table 1, shows the results from the projectivity-based search method, which in this case has explored a total of 231 projected designs.

The couple of factors E and H give the best fit, although a number of three-factor subsets just have $\hat{\sigma}$ values slightly higher. We opt, of course, for the more parsimonious subset in the two factors E and H, which are therefore the active factors. There is thus the opportunity to move on the second stage and analyze with the bootstrap the single projected design in factors E and H. This is a 2^2 full factorial design with a total of $10 \times 3 = 30$ replicates. Table 3 reports, for each term in the full regression model, the mean of the bootstrap distribution, the standard error, and the 95% confidence limits. These results suggest that the main effects of factors E and H, as well as their interaction, are strongly significant. The two analyses in Wu and Hamada (2000) have reached the same conclusions.

Table 2 Top five subsets of lowest $\hat{\sigma}$, thermostat example

One factor	$\hat{\sigma}$	Two factors	$\hat{\sigma}$	Three factors	$\hat{\sigma}$
E	2374.62	E, H	1359.76	E, H, I	1372.58
H	2435.66	E, I	2262.76	C, E, H	1372.73
G	2722.59	E, G	2269.02	E, G, H	1372.73
I	2723.32	C, E	2270.32	F, G, K	1373.27
C	2724.52	C, G	2283.14	C, D, J	1373.33

Table 3 Results from bootstrap, thermostat example

Terms	Mean	Standard error	95% LCL	95% UCL
mean	1695	122	1447	1911
E	−1434	122	−1655	−1186
H	−1329	122	−1548	−1080
EH	1414	122	1165	1632

The third example focuses on data based on an unreplicated 12-run Plackett-Burman design for the study of five factors, denoted by $A - E$, and aims at showing application of the two-stage procedure when $P = 3$ factors turn out to be active in such a plan. The design and the data were extracted by Box and Meyer (1993) from the 2^5 *reactor example* in Box et al. (1978, p. 376), in order to apply their search method for active factors with this partial dataset. This design was also analyzed by Tyssedal and Samset (1997) by the projectivity-based method. It is known from the full analysis that the main effects B, D, E, and the two-factor interactions BD and DE are significant. Both the Box-Meyer method and the projectivity-based method can identify correctly the three active factors B, D, and E. If interest is in bootstrapping to estimate all the potential effects of these active factors, the projected 2^3 design in factors B, D, and E need to be augmented with four additional runs, which are the four unreplicated points in the projection. Due to the availability of the full dataset with 32 runs, we have three possible replicates for each of these four unreplicated points, with a total of 81 potential sets of follow-up runs. With all these sets, 81 possible augmented designs were obtained, each of them being a 2^3 full factorial in factors B, D, and E, replicated twice. A bootstrap analysis, similar to that of the preceding example, was performed on each of these 81 designs, reaching in all cases the same conclusions as in the full analysis. In this constructed example regression could also be applied to the original projected design. More generally, the need for bootstrap and the augmented projection arises in cases with nonstandard data.

5 Conclusion

It is argued that the use of the bootstrap in the analysis of fractionated designs is closely connected with their projective properties. A two-stage approach has been presented for two-level orthogonal designs of projectivity $P = 3$, which are

extremely important in screening experimentation under the precept of factor sparsity. The suggested approach allows to perform a thorough bootstrap analysis after the likely active factors have been appropriately found, and can reveal important interactions in addition to the main effects, but it relies heavily on factor sparsity. When the procedure is inconclusive, however, it can suggest the most appropriate follow-up runs for further study. The value of this approach is especially appreciated in any case where standard regression is not adequate, as evidenced by some of the examples reported.

References

Box, G. E. P., & Hunter, J. S. (1961). The 2^{k-p} fractional factorial designs, part I. *Technometrics*, *3*, 311–351.

Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters*. New York: Wiley.

Box, G. E. P., & Meyer, R. D. (1993). Finding the active factors in fractionated screening experiments. *The Journal of Quality Technology*, *25*, 94–105.

Box, G. E. P., & Tyssedal, J. (1996). Projective properties of certain orthogonal arrays. *Biometrika*, *83*, 950–955.

Box, G. E. P., & Tyssedal, J. (2001). Sixteen run designs of high projectivity for factor screening. *Commun. Statistics–Simulation*, *30*, 217–228.

Cheng, C. S. (1995). Some projection properties of orthogonal arrays. *Annals of Statistics*, *23*, 1223–1233.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge: Cambridge University Press.

Hamada, M., & Wu, C. F. J. (1992). Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, *24*, 130–137.

Jacroux, M. (2007). Main effect designs. In F. Ruggeri, R. S. Kenett, & F. W. Faltin (Eds.), *Encyclopedia of statistics in quality and reliability* (pp. 983–990). Chichester: Wiley.

Kenett, R. S., Rahav, E., & Steinberg, D. M. (2006). Bootstrap analysis of designed experiments. *Quality and Reliability Engineering International*, *22*, 659–667.

Lin, D. K. J., & Draper, N. R. (1992). Projection properties of Plackett and Burman designs. *Technometrics*, *34*, 423–428.

Plackett, R. L., & Burman, J. P. (1946). The design of optimum multifactorial experiments. *Biometrika*, *33*, 305–325.

Tyssedal, J. (2007). Projectivity in experimental designs. In F. Ruggeri, R. S. Kenett, & F. W. Faltin (Eds.), *Encyclopedia of statistics in quality and reliability* (pp. 1516–1520). Chichester: Wiley.

Tyssedal, J., Grinde, H., & Rostad, C. C. (2006). The use of a 12 run Plackett-Burman design in the injection moulding of a technical plastic component. *Quality and Reliability Engineering International*, *22*, 651–657.

Tyssedal, J., & Samset, O. (1997). *Analysis of the 12-run Plackett-Burman design* (Preprint Statistics No. 8/1997). Norway: Norwegian University of Science and Technology.

Wu, C. F. J., & Hamada, M. (2000). *Experiments: Planning, analysis, and parameter design optimization*. New York: Wiley.

CRAGGING Measures of Variable Importance for Data with Hierarchical Structure

Marika Vezzoli and Paola Zuccolotto

Abstract This paper focuses on algorithmic Variable Importance measurement when hierarchically structured data sets are used. Ensemble learning algorithms (such as Random Forest or Gradient Boosting Machine), which are frequently used to assess the Variable Importance, are unsuitable for exploring hierarchical data. For this reason an ensemble learning algorithm called CRAGGING has been recently proposed. The aim of this paper is to introduce the CRAGGING Variable Importance measures, then inspecting how they perform empirically.

1 Introduction

Learning ensemble algorithms, also known as Committee Methods or Model Combiners, are machine learning methods that are implemented within the prediction framework. Given a dataset $(Y, \mathbf{X})$, where Y is a response variable and $\mathbf{X} = \{X_1, X_2, \cdots, X_r, \cdots, X_R\}$ is a set of potential predictors, a learning ensemble mechanism consists in repeatedly fitting the data with a randomized base learner, and then averaging the predictions obtained in each iteration (see Breiman 2001; Friedman and Popescu 2003). Learning ensembles have proven to be accurate in terms of predictability. In addition, they partially overcome the well-known problem of the black-box, as they allow the computation of importance measures of the predictors. Since empirical analysis often deals with very large data sets, where important predictors are hidden among a great number of uninformative covariates, this feature is highly desirable.

When we use decision trees as base learner, we are dealing with the so called tree-based learning ensembles, which are the focus of this paper. The most important VI measurement methods in the context of decision trees and tree-based learning ensembles have been proposed in Breiman et al. (1984), Breiman (2001) and Friedman (2001). The most common approaches are based on two measures: the *Mean Decrease in Accuracy* (M1 henceforth) and the *Total Decrease in Node Impurity* (M2 henceforth).

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_44,

The basic idea behind M1 is that a random permutation of the values of the predictor variable X_r is supposed to mimic the absence of the variable from the model. The difference in the prediction accuracy before and after permuting the predictor variable is used as a VI measure.

M2 derives from the principle of impurity reduction within decision trees. Briefly, the reductions in the heterogeneity of Y computed in each node by a given variable X_r are summed up over all the nodes with X_r as splitting variable. Despite its large use, this measure has been shown to be biased when predictor variables have different numbers of categories or measurement scales (Shin and Tsai 2004; Sandri and Zuccolotto 2008; Sandri and Zuccolotto 2010). On this issue, some authors have recently proposed methods aimed at eliminating the bias of the M2 measure (see Kim and Loh 2001; Sandri and Zuccolotto 2008; Sandri and Zuccolotto 2010).

Although mainly formalized in the context of Random Forest (Breiman 2001) and Gradient Boosting Machine (Friedman 2001), the basic idea supporting these two measures can be conceptually applied to every tree-based learning ensemble.

The aim of this paper is to introduce the measures M1 and M2 in the context of the novel ensemble learning called CRAGGING (**CR**oss-validation **AGG**regat**ING**), introduced in Vezzoli (2007) and Vezzoli and Stone (2007) in order to fit hierarchical data. Mainly focusing on the M1 measure, we modify the underlying permutation mechanism in a way which is consistent with the specific structure of the data.

The paper is organized as follows: in Sect. 2 there is a brief formalization of the CRAGGING algorithm while in Sect. 3 we propose the formulation of its VI measures. In Sect. 4 we compare the performance of Logistic Stepwise Regression, Lasso Logistic Regression, Random Forest and CRAGGING on the same data set. Some final remarks conclude the paper.

2 A Brief Description of CRAGGING

Let $(Y, \mathbf{X})$ be a data set with N observations, where Y is the response variable and $\mathbf{X}$ is the matrix of the R predictors. We assume that the observations are divided into J groups according to a categorical variable F. Each group is composed by n_j observations and $N = \sum_{j=1}^{J} n_j$. Let us denote with $\mathscr{L} = \{1, 2, \cdots, J\}$ the set of groups and with $\mathbf{x}_{ji} = (x_{1ji}, x_{2ji}, \cdots, x_{rji}, \cdots, x_{Rji})$ the vector of predictors for i-th subject of group j where $j \in \mathscr{L}$ and $i = 1, 2, \cdots, n_j$. CRAGGING works as follows. Firstly, the set $\mathscr{L}$ is randomly partitioned in V subsets denoted by $\mathscr{L}_v$ with $v = 1, \cdots, V$, each one containing J_v groups. For each v, let $\mathscr{L}_v^c$ be the complementary set of $\mathscr{L}_v$, containing J_v^c groups. In addition, let $\mathscr{L}_{v\setminus\ell}^c$ be the set obtained by removing the ℓ-th group from $\mathscr{L}_v^c$ ($\ell \in \mathscr{L}_v^c$ and $\mathscr{L}_{v\setminus\ell}^c \bigcup \ell = \mathscr{L}_v^c$). Secondly, for a fixed α, for each $\mathscr{L}_v$ and for each $\ell \in \mathscr{L}_v^c$ let

$$\hat{f}_{\boldsymbol{\alpha}, \mathscr{L}_{v\setminus\ell}^c}(\cdot) \tag{1}$$

be the prediction function of a single tree (base learner) trained on data $\{y_{ji}, \mathbf{x}_{ji}\}_{j \in \mathscr{L}^c_{v\setminus\ell},\, i=1,\cdots,n_j}$ and pruned with cost-complexity parameter α. The corresponding prediction for the observations not used to grow the tree (test set) is given by

$$\hat{y}_{ji,\alpha\ell} = \hat{f}_{\alpha,\mathscr{L}^c_{v\setminus\ell}}(\mathbf{x}_{ji}), \text{ with } j \in \mathscr{L}_v, \text{ and } i = 1, 2, \cdots, n_j. \tag{2}$$

According to the learning ensemble's philosophy, an aggregated prediction over the groups contained within the test set $\{y_{ji}; \mathbf{x}_{ji}\}_{j \in \mathscr{L}_v,\, i=1,\cdots,n_j}$ is obtained by the average of functions (2):

$$\hat{y}_{ji,\alpha} = \frac{1}{J^c_v} \sum_{\ell \in \mathscr{L}^c_v} \hat{f}_{\alpha,\mathscr{L}^c_{v\setminus\ell}}(\mathbf{x}_{ji}) \text{ with } j \in \mathscr{L}_v \text{ and } i = 1, 2, \cdots, n_j. \tag{3}$$

Thirdly, the procedure is repeated for different values of α and finally the algorithm chooses the optimal tuning parameter α^*. Such parameter corresponds to that value for which the out of sample error estimation over all $\mathscr{L}_v$ is minimized:

$$\alpha^* = \arg\min_{\alpha} L(y_{ji}, \hat{y}_{ji,\alpha}) \text{ with } j \in \mathscr{L}, \; i = 1, 2, \cdots, n_j \tag{4}$$

where $L(\cdot)$ is a generic loss function. The CRAGGING predictions are given by

$$\tilde{y}^{\text{crag}}_{ji} = \hat{y}_{ji,\alpha^*} \text{ with} j \in \mathscr{L}, \; i = 1, 2, \cdots, n_j.$$

It is worth noting that the loss function $L_{\alpha^*} = L(y_{ji}, \hat{y}_{ji,\alpha^*})$ is able to measure the generalization error of the algorithm, based on the predictions for a generic subject i computed using trees grown with training sets not containing that subject. This is somewhat similar to the out-of-bag estimation of prediction error, which is used, for example, by the Random Forests. However, the mechanism through which a sub sample is removed from the training set at each iteration differs from bagging, as it derives from a cross-validation rotation. Thus we introduce the term *out-of-crag* (OOC) estimation of prediction error. Similarly to what happens with bagging, the OOC prediction error tends to overestimate the real generalization error of the algorithm, because predictions are computed using only a small subset of the total trees composing the ensemble learning.

3 VI Measures with CRAGGING

In this section we introduce the M1 and M2 measures in the context of CRAGGING.

M1$_{CR}$: at each tree of CRAGGING in correspondence of α^*, all the values of the r-th variable are randomly permuted and new predictions are obtained with this new

data set $(Y, \mathbf{X})_r$. Hence, we compute a new loss function $L_{\alpha^*,r}$ and we compare it with $L_{\alpha^*} = L(y_{ji}, \hat{y}_{ji,\alpha^*})$.

To be coherent with the idea of perturbing the training set without destroying the structure of the data, we randomize the values of X_r conditionally to the J groups of the data set. In other words, for each variable X_r a permutation $p = \{p_1, p_2, \cdots, p_J\}$ of the sequence $\{1, 2, \cdots, J\}$ is randomly selected. The values of X_r are randomized in the data set according to the following rule:

$$\{x_{rji}\}_{j \in \mathscr{L},\, i=1,2,\cdots,n_j} = s(\mathbf{x}_{rp_j}), \tag{5}$$

where $s(\cdot)$ denotes a sampling with replacement from a set of values and $\mathbf{x}_{rp_j} = \{x_{rp_j i}\}_{i=1,\cdots,n_{p_j}}$. This way of randomizing the values of X_r should be particulary useful if $f(X_r|j_1) \neq f(X_r|j_2)$ for all $j_1 \neq j_2$, as frequently happens in the application domain of CRAGGING. The procedure of sampling is repeated k times and the M1 measure for the r-th variable is given by the following average on k:

$$M1_r = av_k(L_{\alpha^*,r} - L_{\alpha^*}).$$

M2$_{CR}$: at each tree of CRAGGING the heterogeneity reductions due to variable X_r over the set of nonterminal nodes are summed up and the importance of variable X_r is computed averaging the results over all the trees of the ensemble. Formally, let d_{rg}^t be the decrease in the heterogeneity index due to X_r at the nonterminal node $g \in G$ of the t-th tree ($t = 1, \cdots, T$). The VI of r-th variable over all the trees is:

$$\widehat{VI}_{X_r} = \frac{1}{T} \sum_{t=1}^{T} \sum_{g \in G} d_{rg}^t I_{rg}^t \tag{6}$$

where I_{rg}^t is the indicator function which equals 1 if the r-th variable is used to split node g and 0 otherwise.

4 Case Study

Using data from Standard & Poor's and International Monetary Fund we analyzed sovereign defaults over the period 1975–2002 for 66 emerging countries with a dichotomous response variable denoting crisis (1) and not crisis (0) events. The data set is the same used in Savona and Vezzoli (2008) and includes 22 potential predictors, both quantitative and qualitative, comprising measures of external debt, public debt, solvency and liquidity, financial variables, real sector and institutional factors. The predictors are lagged one year in order to focus on default predictability.

In the empirical analysis we compared the performance of different predicting methods, using the Area Over the Roc Curve (AOC henceforth) to measure the model's inaccuracy. In more depth:

- We fitted the Logistic Stepwise Regression (LSR) on a training set containing 63.2% of the observations.[1] The accuracy of predictions was assessed on the remaining data. We randomly repeated the procedure 1,000 times, for each model we computed the AOC, then using the average of the areas computed at each interaction as a global measure of accuracy (**AOC = 0.2444**).
- We used the same procedure by fitting the Lasso Logistic Regression (LASSO), a shrinkage method based on optimal constraint on coefficient estimates in order to select the best predictors among possible competing candidates (Tibshirani, 1996) (**AOC = 0.2294**).
- In order to compare the Random Forest (RF) with the CRAGGING, we have grown a forest of 660 regression trees[2] (**AOC = 0.3944**). In addition, since Breiman (2001) proved that the RF performance can be improved by increasing the number of trees, we have grown a forest of 5,000 trees (**AOC = 0.2152**).
- Finally we carried out the CRAGGING by randomly splitting the $J = 66$ groups in $V = 11$ subsets, each one containing $J_v = 6$ countries. For a fixed α and for each training set with $J_v^c = 60$ groups, we computed (1) and the corresponding predictions in the test set (2) and (3). The procedure is repeated for different values of α until the optimal value α^* is selected[3] (**AOC = 0.2406**).

The results of the empirical analysis show that LSR, LASSO and CRAGGING perform quite similarly. On its hand, RF with 660 trees underperforms the CRAGGING with the same number of trees, but when 5,000 trees are grown, the AOC of RF decreases by around 45%, achieving the best result obtained in this case study.[4] As stated before, we have to keep in mind that CRAGGING overestimates the current AOC because for each subject i, the prediction $\tilde{y}_{ji}^{\text{crag}}$ is obtained by averaging only $J_v^c = 60$ trees out of the 660 trees grown.

Before analyzing the VI measures, let us consider Fig. 1, which reports the cross section CRAGGING probabilities of default, fitted by a Gaussian kernel smoothing non parametric regression, compared with the real GDP growth over the period 1975–2002. Default probabilities exhibit a cyclical pattern, which appears strongly correlated with the business cycle as proxied by the GDP growth.

Figure 2 shows the VI measure for LSR and LASSO, computed for each variable X_r as the fraction of iterations where X_r has been selected among the relevant predictors (AT measure, Austin and Tu 2004). The VI measures for RF with 5,000 trees and CRAGGING are displayed in Figs. 3 and 4, respectively.

[1] This is the same percentage of instances that the default setting in the R package `randomForest` suggests for each bootstrap training set used in the out-of-bag estimates.

[2] This is the same number of trees composing CRAGGING that is $V \times J_v^c = 11 \times 60$.

[3] In our case the value of α^* minimizing the AOC is $\alpha^* = 0.6491$ which means a low penalty on the number of leaves.

[4] It is well-known that as the number of base learners increases, ensemble learning algorithms improve their performance. Further research is currently devoted to explore this issue also for the CRAGGING, where the number of trees could be increased by repeating M times the algorithm and then averaging the results.

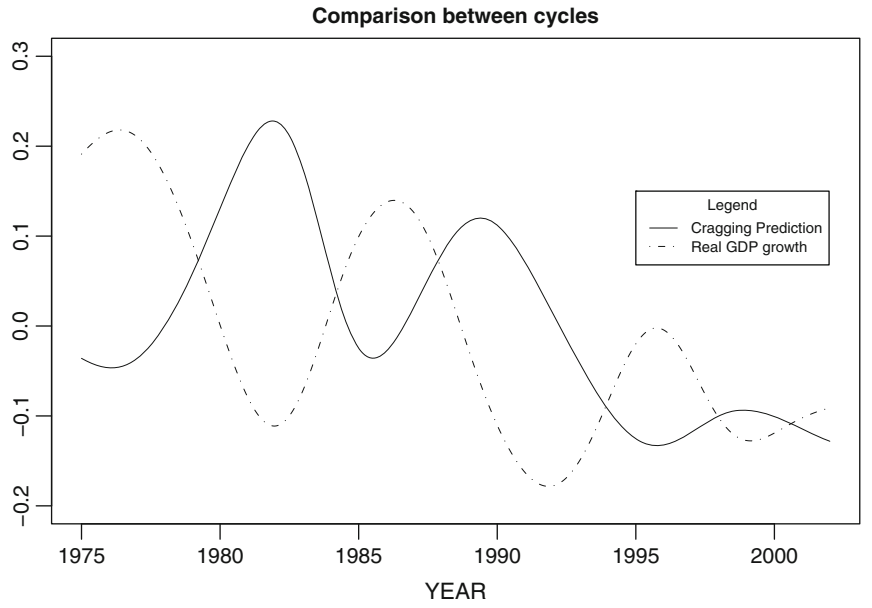


Fig. 1 Sovereign default and GDP cycles over the period 1975–2002

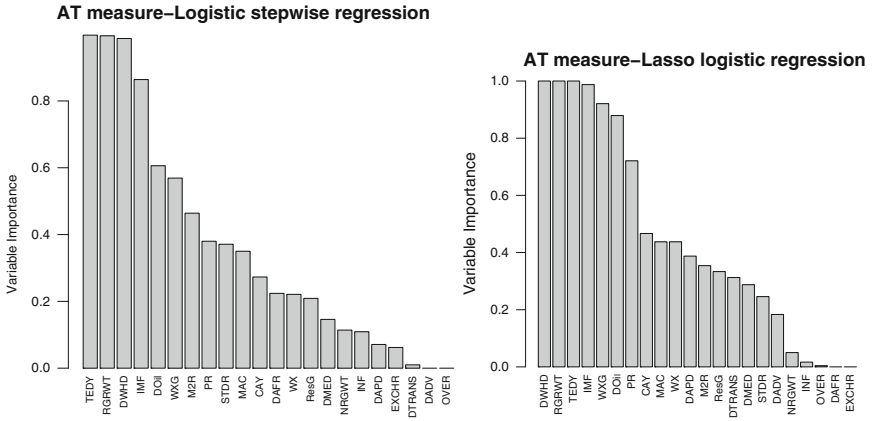


Fig. 2 VI measure for LSR and LASSO

In order to inspect the concordance among the rankings of the predictors based on the different VI measures, we also computed the Spearman rank correlation. The results are in Table 1.

We observe a concordance among the rankings obtained by means of $\mathbf{M1}_{RF}$, $\mathbf{M2}_{RF}$ and $\mathbf{M2}_{CR}$. On their hand, LSR and LASSO exhibit low concordance when compared against other methods, and high concordance when compared each other. Note, in particular, that INF (inflation) and NRGWT (nominal GDP growth) are

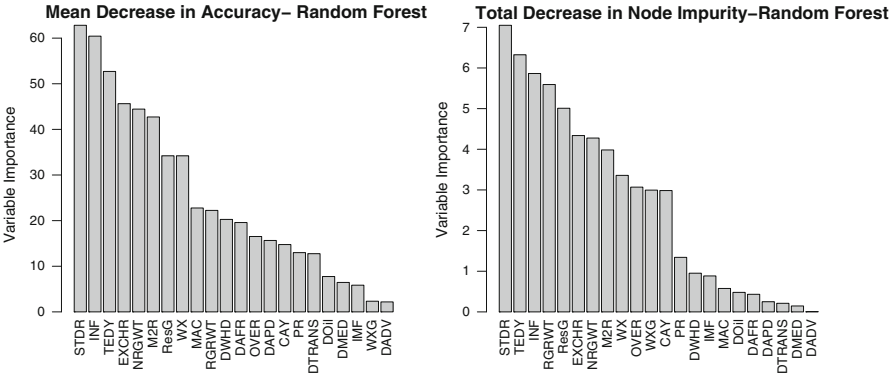


Fig. 3 VI measures for RF (5,000 trees)

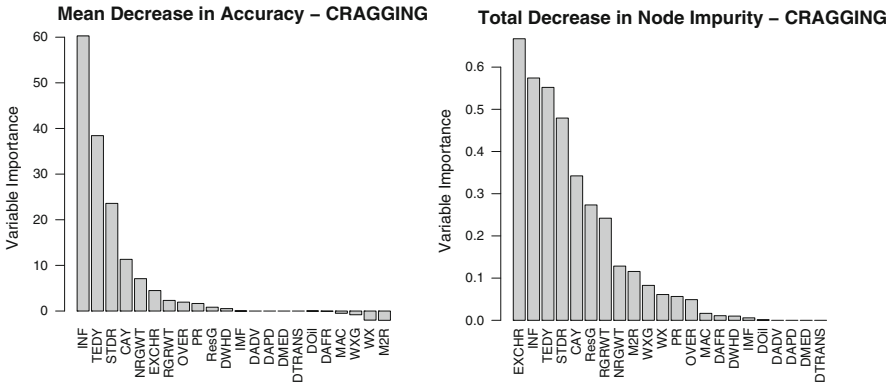


Fig. 4 VI measures for CRAGGING

Table 1 Spearman coefficients between the variable rankings defined by VI measures

	LSR	LASSO	$\mathrm{M1}_{RF}$	$\mathrm{M2}_{RF}$	$\mathrm{M1}_{CR}$	$\mathrm{M2}_{CR}$
LSR	1					
LASSO	0.8261	1				
$\mathbf{M1}_{RF}$	0.0423	−0.2547	1			
$\mathbf{M2}_{RF}$	0.2637	0.0051	0.8080	1		
$\mathbf{M1}_{CR}$	−0.0390	−0.1214	0.4647	0.5844	1	
$\mathbf{M2}_{CR}$	0.1711	−0.1011	0.7549	0.9164	0.6126	1

considered as not much informative by LSR and LASSO, while RF and CRAGGING address these two variables as the "key" factors in predicting sovereign default, as proven also by many economic studies. This is probably due to the multicollinearity between INF and NRGWT ($\rho = 0.9639$), which could have caused some problems when using LSR and LASSO. Looking at $\mathbf{M1}_{CR}$, we observe that it exhibits a moderately low concordance with other measures. This could be due

to the specific randomization criterion introduced in Sect. 3. As a result, we should perform simulation studies in order to assess its reliability, as well as investigate if the problem of bias affecting the M2 measure also arises for the CRAGGING algorithm, which uses pruned trees. In fact, a recent study has proven that the bias is generated by uninformative splits which can be removed by pruning (Sandri and Zuccolotto 2010). Such an issue deserves a dedicated study that we leave to our future research agenda.

References

Austin, P. C., & Tu, J. V. (2004). Bootstrap methods for developing predictive models. *The American Statisticians*, *58*(2), 131–137.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Breiman, L., Friedman, J. H., Olshen, R., & Stone, C.J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth and Brooks/Cole.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232.

Friedman, J. H., & Popescu, B. E. (2003). *Importance sampled learning ensembles*. Stanford University, Department of Statistics, Technical Report.

Kim, H., & Loh, W. Y. (2001). Classification trees with unbiased multiways splits. *Journal of the American Statistical Association*, *96*, 589–604.

Sandri, M., & Zuccolotto, P. (2008). A bias correction algorithm for the Gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics*, *17*, 1–18.

Sandri, M., & Zuccolotto, P. (2010) Analysis and correction of bias in Total Decrease in Node Impurity measures for tree-based algorithms. *Statistics and Computing, 20*(4), 393–407.

Savona, R., & Vezzoli, M. (2008). *Multidimensional distance to collapse point and sovereign default prediction*. Carefin Working Paper, 12/08, Milano.

Shin, Y., & Tsai, H. (2004). Variable selection bias in regression trees with constant fits. *Computational Statistics and Data Analysis*, *45*(3), 595–607.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, *25*, 267–288.

Vezzoli, M. (2007). *Recent advances on Classification and Regression Trees*, Unpublished PhD Thesis, University of Milano Bicocca.

Vezzoli, M., & Stone, C. J. (2007). CRAGGING. In *Book of short papers CLADAG 2007* (pp. 363–366). EUM.

Regression Trees with Moderating Effects

Gianfranco Giordano and Massimo Aria

Abstract This paper proposes a regression tree methodology that considers the relationships among variables belonging to different levels of a data matrix which is characterized by a hierarchical structure. In such way we consider two kinds of partitioning criteria dealing with non parametric regression analysis. The proposal is based on a generalization of Classification and Regression Trees algorithm (CART) that considers a different role played by moderating variables. In the work are showed some applications on real and simulated dataset to compare the proposal with classical approaches.

1 Introduction

Within several research fields (sociology, economics, demography and health), it is likely to deal with hierarchical structure phenomenon, with multi-level data: individual, familiar, territorial and social. In such circumstances it is necessary to proceed with the analysis of the relation between individuals and the society, where naturally, can be observed at different hierarchical levels, and variables may be defined at each level (Leeuw and Meijer 2008). This leads to research into the interaction between variables characterizing individuals and variables characterizing groups. We define the measurement of this interaction as moderating effect.

The most of the applications, for which the classical approaches are used, consider the noticed fundamental units as belonging to one only set and deriving from the same population. Historically, multilevel problems led to analysis approaches that moved all variables by aggregation or disaggregation to one single levels of interest followed by *ordinary multiple regression*, *analysis of variance*, or some other standard analysis method (Hox 2002). However, analyzing variables from different levels at one single common level is inadequate because it leads to ignore the stratifications and the hierarchies.

Multilevel models are statistical methodologies that are more adequate to better extract the information of typical hierarchical structures. Those take into account the presence of relations among variables belonging to either the same level or different ones, considering the net effect of the units and their interactions.

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_45,

Among the main limits of such an approach there is that multilevel model so far require that the grouping criterion is clear, the variables must be assigned unequivocally to their appropriate level and, this models have a wide range of theoretical assumptions and an equal wide range of specifications problems for the auxiliary theory (Snijders and Bosker 1999). Moreover, when at each level there are several variables, there is a huge amount of possible cross level interactions making the parameter estimation and interpretation very hard.

The main objective of this work is to analyze such type of matrices of data through an innovative regression tree methodology, trying to stress the potentialities and the advantages of such non parametric techniques.

2 Regression Trees

As with all regression techniques we assume the existence of a single output (response) variable and one or more input (predictor) variables. The output variable is numerical. The general regression tree methodology allows input variables to be a mixture of continuous and categorical variables (Breiman et al. 1984). Regression tree is built through a process known as *binary recursive partitioning*. This is an iterative process of splitting the data into partitions, and then splitting it up further on each of the branches. Initially all of the records in training set (the preclassified records that are used to determine the structure of the tree) are together in one group (root node). The algorithm then tries breaking up the data, using every possible binary split on every field. The algorithm chooses the split that partitions the data into two parts such that it minimizes the impurity in the children nodes. This splitting or partitioning is then applied to each of the new branches. The process continues until each node reaches a user-specified stopping rule and becomes a terminal node.

Summarizing, tree based methods involve the following steps:

- the definition of a splitting (partitioning) criterion;
- the definition of response value to the terminal nodes;
- tree pruning, aimed at simplifying the tree structure, and tree selection, aimed at selecting the final decision tree for decisional purposes.

2.1 CART Partitioning Criteria

Let Y, X be a multivariate variable where X is a set of K categorical or numerical predictors $(X_1 \ldots, X_k \ldots, X_K)$ and Y is the response variable.

Let S, the finite set of possible splits (dummy variables) generated from all possible binarization of X predictors.

Once that, at a given node, the set of binary questions has been created, some criterion which guides the search in order to choose the best one to split the node

is needed. Therefore we split each node so that each descendant is more homogeneous than the data in the parent node. To reach this aims, we need a measure of homogeneity to be evaluated by means of a splitting criterion.

In the Classification and Regression Trees methodology (CART, Breiman et al. 1984), the best binary cut of a *parent node* is find minimizing the impurity of response in the left and right children node respectively.

This is equivalent to maximize the decrease of impurity due to the split s

$$\Delta i(t,s) = i(t) - [i(t_l) \cdot p_l(t) + i(t_r) \cdot p_r(t)] \tag{1}$$

where $i(t)$ is the impurity index in the parent node t, $i(t_l)$ and $i(t_r)$ are the impurity in the left and right children node, and p_l and p_r are the proportion of cases in each node (Mola and Siciliano 1997).

The measure $\Delta i(t,s)$ can be seen as the gain, in terms of the response purity, obtained by the split $s \in S$.

In regression analysis, the impurity is measured by the variance of response variable in the node t weighted by the proportion of cases in the same node

$$i_{CART}(t) = \frac{\sum_{i \in t}(y_i - \overline{y}_t)^2}{N_t} \frac{N_t}{N} = \frac{TSS(Y(t))}{N} \tag{2}$$

where N_t is the number of cases in the node t and N is the total sample size, and $TSS(Y(t))$ is the total sum of squares of Y at node t. Among the typical issues of the CART, complex relationships in the data are often neglected, because they are based on tree structures that comes to partitioning using, step by step, the split generated by X predictor not considering the presence of moderating variables. Moreover pattern and levels (stratifications, hierarchy, etc.) in data are ignored (Siciliano et al., 2004).

3 Regression Trees with Moderating Effects

The proposed methodology, namely *Regression Trees with Moderating Effects*, is based on the generalization of CART partitioning criteria, through the definition of two splitting algorithms that take into account the main moderating effect of the Z variable with respect to the prediction of X over Y. The goal of such an approach is to define a recursive partitioning algorithm that *identifies the best final conditioned partition with one or more moderating variables expression of the stratification hierarchy* (Siciliano et al., 2007, Tutore et al., 2007).

The relation between the Z moderating variable and the response variable Y indirectly conditions the relation of X over Y, acting on the strength of their link. In order to face this situation, we proposed two alternative partitioning criterion which measure the moderating effect combining the CART classical impurity index with a multiplicative or an additive contribution of variable Z.

3.1 Multiplicative Partitioning Criterion

The first proposal is based on intra-class correlation (ICC) index to consider the role played by Z in the explanation of Y. In particular, we consider a class of ICC measures which have the following properties:

- the index is null when the moderating effect is absent;
- the index increases as the moderating effect grows;
- the index is equal to 1 in case of the moderating effect is maximum.

An index with such properties is a well known Intraclass Correlation Coefficient proposed by Donner (1986)

$$\rho_{ICC} = \frac{\text{var}\,(between\ classes)}{\text{var}\,(between\ classes) + \text{var}\,(residual)} \tag{3}$$

where the word 'class' means a level of a population layer with the presence of a moderating variable, expression of the data information hierarchy.

The intra-class correlation is used to estimate the correlation of one variable between two members within a group, for instance between two children of one family. In other words, the intraclass correlation gives the proportion of variance attributable to between group differences.

Let t a generic node of the tree, we define the impurity measure with multiplicative effect as

$$i_m(t) = \left[\frac{TSS(Y_t)}{N}\right] \cdot [1 - \rho_{Y|Z}(t)] \tag{4}$$

where

- $TSS(Y(t))/N$ is the classical CART impurity measure;
- $\rho_{Y|Z}$ is the intraclass correlation coefficient of Y given the stratifying defined by Z at node t.

Moreover we define the decrease of impurity as (1) and identify the best split $s^* \in S$ as

$$\Delta i\,(t\,|s*) = \max! \tag{5}$$

3.2 Additive Partitioning Criterion

A second proposal to treat moderating effects is represented by the definition of a impurity measure that considers an additive effect of Z on causal link between response and predictors

$$Prediction\ of\ Y = effect\ of\ X + moderating\ effect\ of\ Z$$

The impurity in a node t is:

$$i_a(t) = \left[\frac{TSS(Y_t)}{N} + \sum_{h=1}^{H} \frac{WSS(Y_t \mid Z_h)}{g_h(t)} \right] \tag{6}$$

where

- $WSS(Y_t \mid Z_h)$ is the within sum of squares of the Y at t node conditioned at h group of the Z;
- $g_h(t)$ are the degrees of freedom for each group at node t.

Equation (6) defines the impurity measure at node t as the additive combination of X contribution to the Y prediction, taking into account, at the same time, the effect of split on the conditioned distribution of Y respect to Z.

Following the previous approach, we maximize the decrease of impurity at every generic node t (5). It is possible to demonstrate, for both the proposed criterion, that, in absence of moderating effects, RTME can be consider as a CART generalization. In fact, when there is not influence of Z, the different partitioning criteria coincide:

$$i_m(t) = \left[\frac{TSS(Y_t)}{N} \right] * \left[1 - \rho_{Y|Z}(t) \right] \equiv i(t)_{CART} \tag{7}$$

and

$$i_a(t) = \left[\frac{TSS(Y_t)}{N} + \sum_{h=1}^{H} \frac{WSS(Y_t \mid Z_h)}{g_h(t)} \right] \equiv 2 \cdot i(t)_{CART} \tag{8}$$

and the resulting trees produce the same partition.

In both approaches, the search of the best split s^* consist on the identification of the best binary partition as a compromise between the X prediction strength and the ability to express the moderating effect of Z. This compromise can be considered in an additive or multiplicative way depending on the typology of moderating link.

4 Comparison Study and Concluding Remarks

The performance of the proposed method based on *additive* and *multiplicative* criterion has been evaluated in several comparative analysis (Giordano and Remmerswaal 2009). In the following are shown the results of the non-parametric approach, using the additive and multiplicative criteria, that have been compared with the results of the multilevel model. Two simulated datasets (Table 2) and three well-known real datasets (Table 1) are considered.

The results are shown in Tables 3 and 4. In order to compare the overall accuracy between the observed and the predicted variables of the models, we report some accuracy measures:

- The *classical Goodness of Fit* (GoF overall)
- The *Moderating Goodness of Fit* (GoF moderating) that shows the accuracy gain, in terms of within sum of squares (WSS) of $Y|z$

Table 1 Description of real datasets

Dataset name	Sugar cane	Pulse rate	ILE authority
Description	*This data gives sugar cane yields for each paddock in the North Queensland for the 1997 sugar cane season*	*Pulse Rates before and after Exercise. The pulse rates and other physiological and lifestyle data are given in the data*	*Data consisting of examination records from 140 secondary schools in different years*
Source	Denman, N., and Gregory, D. (1998)	R. J. Wilson, Univ. of Queensland (1998)	ILEA Research and Statistics (1987)
Response	Cane Quality	Relative Pulse Difference	Student's Score
Moderating	Districts	Excercise Type	Student's Country
N. of Predictors	*24*	*8*	*8*

Table 2 Description of simulated datasets

Dataset name	Simulation 1	Simulation 2
Predictors	*Predictors have been generated by different random distributions (Discrete Unif., Continue Unif., Multnomial, Normal) five predictors*	*Predictors have been generated by different random distributions (Discrete Unif., Continue Unif., Multnomial, Normal) eight predictors*
Response	linear link with predictors four groups	nonlinear link with predictors five groups
Moderating	two levels of influence	five levels of influence
N. of cases	*1,000*	*5,000*

Table 3 Comparision among multilevel methods (Real datasets)

Dataset	Partitioning criteria	Number of nodes	Overall GoF	Moderating GoF	Overall AC ratio	Moderating AC ratio
Sugar Cane	*Multilevel Analysis*	–	0,2785	0,1575	–	–
	CART impurity	96	0,4593	0,4219	4,7844	4,3948
mod.effect	*RTME Additive impurity*	79	0,4464	0,4302	5,6506	5,4456
0,1192	*RTME Multiplicative impurity*	69	0,4123	0,4672	5,9754	6,7710
Pulse	*Multilevel Analysis*	–	0,7380	0,7306	–	–
	CART impurity	23	0,9157	0,9542	39,8130	41,4870
mod.effect	*RTME Additive impurity*	21	0,9221	0,9304	43,9095	44,3048
0,0110	*RTME Multiplicative impurity*	21	0,8943	0,9467	42,5857	45,0810
ILE	*Multilevel Analysis*	–	0,3395	0,0925	–	–
	CART impurity	105	0,1378	0,1578	1,3124	1,5029
mod.effect	*RTME Additive impurity*	109	0,2655	0,1712	2,4358	1,5706
0,2705	*RTME Multiplicative impurity*	104	0,2702	0,1907	2,5981	1,8337

Table 4 Comparision among multilevel methods (Simulated datasets)

Dataset	Partitioning criteria	Number of nodes	Gof overall	GoF moderating	Overall AC ratio	Moderating AC ratio
Sim 1	*Multilevel Analysis*	–	0,3492	0,6759	–	–
	CART impurity	26	0,4414	0,5037	0,01698	0,01937
mod.effect	*RTME Additive impurity*	24	0,4690	0,5420	0,01954	0,02258
0,0010	*RTME Multiplicative impurity*	20	0,4414	0,5203	0,02207	0,02602
Sim 1	*Multilevel Analysis*	–	0,9553	0,5161	–	–
	CART impurity	26	0,4414	0,1366	0,01698	0,00525
mod.effect	*RTME Additive impurity*	20	0,7009	0,1410	0,03505	0,00705
0,9541	*RTME Multiplicative impurity*	20	0,6114	0,3525	0,03057	0,01763
Sim 2	*Multilevel Analysis*	–	0,1217	0,1193	–	–
	CART impurity	31	0,1713	0,1879	0,00553	0,00606
mod.effect	*RTME Additive impurity*	25	0,1755	0,1900	0,00702	0,00760
0,0001	*RTME Multiplicative impurity*	22	0,1679	0,1900	0,00763	0,00864
Sim 2	*Multilevel Analysis*	–	0,4189	0,0747	–	–
	CART impurity	31	0,1713	0,1459	0,00553	0,00471
mod.effect	*RTME Additive impurity*	19	0,3277	0,1864	0,01725	0,00981
0,3690	*RTME Multiplicative impurity*	20	0,3035	0,1523	0,01518	0,00762
Sim 2	*Multilevel Analysis*	–	0,5311	0,0653	–	–
	CART impurity	31	0,1713	0,1506	0,00553	0,00486
mod.effect	*RTME Additive impurity*	28	0,3774	0,1868	0,01348	0,00667
0,4956	*RTME Multiplicative impurity*	24	0,3624	0,1749	0,01510	0,00729
Sim 2	*Multilevel Analysis*	–	0,7258	0,0706	–	–
	CART impurity	31	0,1713	0,1439	0,00553	0,00464
mod.effect	*RTME Additive impurity*	21	0,4543	0,1901	0,02163	0,00905
0,7024	*RTME Multiplicative impurity*	19	0,4493	0,1965	0,02365	0,01034
Sim 2	*Multilevel Analysis*	–	0,9420	0,0079	–	–
	CART impurity	31	0,1713	0,0333	0,00553	0,00107
mod.effect	*RTME Additive impurity*	33	0,5572	0,0381	0,01688	0,00115
0,9414	*RTME Multiplicative impurity*	34	0,5187	0,1187	0,01526	0,00349

$$GoF_{\text{mod}} = 1 - \left(\sum_{j=1}^{J} \left[\frac{\sum_{i=1}^{n_j} (\hat{Y}_{ij} - Y_{ij})^2}{df_j} \right] \cdot \frac{n_j}{N} \Big/ \sum_{j=1}^{J} \left[\frac{\sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{.j})^2}{n_j - 1} \right] \cdot \frac{n_j}{N} \right) \tag{9}$$

Where i indicates a generic individual and j indicates a generic group. To compare trees with different number of terminal nodes, we define the *accuracy ratio* (AC) as the ratio between the GoF measure and the model complexity (size of tree). The number of terminal nodes indicate the complexity of the tree. Higher values of AC ratio mean better ability of the tree method to explain the relationship information with a small structure.

According to the GoF in Tables 3 and 4, the RTME algorithms are more accurate than the CART procedure. Although, the multilevel procedure is in overall

terms still more accurate in small dataset with suitable distributional and functional hypothesis. One of the achievements of this study shows that it is possible to use a non-parametric method for the analysis of arrays of data with hierarchical structure, overcoming the limits of the CART methodology. When in presence of situations where the functional and distributional assumptions of the multilevel model are not verified, or when its estimation algorithm does not converge, it is possible to use this technique to provide a viable and feasible alternative. Looking to the trade-off between the complexity and accuracy, the moderating GoF is larger for the non-parametric approaches in comparison to multilevel analysis. This is an important remark according to the task to better explain the relationship among response and predictors at each level of hierarchy defined in the population. Starting from this point of view, when data are characterized by a linear relationship among variables, the multiplicative impurity measure for RTME is the best method to describe a dataset with moderating effects. On the contrary, the additive criterion is the best choice in presence of a non linear effect.

References

Breiman, L., Friedman, J. H., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Belmont, California: Wadsworth.

de Leeuw, J., & Meijer, E. (2008). Handbook of multilevel analysis. New York: Springer.

Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review*, *54*(1), 67–82.

Giordano, G., & Remmerswaal, R. (2009). *Non-parametric regression model for a hierarchical data-structure: A comparison with the classical approaches*. Seventh scientific meeting of the classification and data analysis group of the Italian statistical society. Book of short papers (Catania, September 09–11, 2009), pp. 259–262.

Hox, J. J. (2002). *Multilevel analysis, techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mola, F., & Siciliano, R. (1997). A fast splitting procedure for classification and regression trees. *Statistics and Computing*, *7*, 208–216.

Siciliano, R., Aria, M., & Conversano, C. (2004). Harvesting trees: Methods, software and applications. In *Proceedings in computational statistics: 16th Symposium of IASC Held in Prague, August 23–27, 2004* (COMPSTAT2004), Electronical Edition (CD). Heidelberg: Physica-Verlag.

Siciliano, R., Tutore, V. A., & Aria, M. (2007). *3Way Trees, Invited paper in Proceedings of Classification and Data Analysis Group (CLADAG 2007)*, Edizioni Universit di Macerata, September 12–14, Macerata.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis*. An introduction to basic and advanced mutilevel modeling. London: SAGE Publications.

Tutore, V. A., Siciliano, R., & Aria, M. (2007). Conditional classification trees using instrumental variables. In *Proceedings of the 7th IDA2007 Conference* (Ljubljana, 6–8 September, 2007), Lecture Notes in Computer Science Series of Springer.

Data Mining for Longitudinal Data with Different Treatments

Mouna Akacha, Thaís C.O. Fonseca, and Silvia Liverani

Abstract The CLAssification and Data Analysis Group (CLADAG) of the Italian Statistical Society recently organised a competition, the 'Young Researcher Data Mining Prize' sponsored by the SAS Institute. This paper was the winning entry and in it we detail our approach to the problem proposed and our results. The main methods used are linear regression, mixture models, Bayesian autoregressive and Bayesian dynamic models.

1 Introduction

Recently the CLAssification and Data Analysis Group (CLADAG) of the Italian Statistical Society organised a competition on a data mining problem. Given the sales of nine products over seven time periods, five structural variables and a marketing campaign treatment for 4,517 sales points, the competitors are asked to evaluate the marketing campaign impact on the economic return in the first time period, and forecast the economic return for the seventh time period.

The first question may be seen as a problem studied in regression analysis, whilst the second problem is widely studied in time series forecasting. However, the presence of several covariates with non-linear and co-dependent features requires both questions to be addressed with ad hoc methods. The main statistical method we use to address the first question is a mixture of regression models, while we fit autoregressive and dynamic models for the forecasting problem.

This paper is organised as follows. In Sect. 2 we describe the data and perform exploratory tests and analysis. In Sect. 3 we detail the mixture model and our results to answer the first question asked by the organisers. In Sect. 4 we introduce the autoregressive and dynamic models for prediction and present our results for the second question.

An extended version of this paper is available as a CRiSM Working Paper (Akacha et al. 2009). Throughout this paper we will refer to the extended paper for more details.

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_46,

2 The Data

The data provided in this competition was collected by sales points over seven time periods. The outcome variable y_{it} is an unknown function of the income of the sales point i during time period t, with $t = 1, \ldots, 7$ and $i = 1, \ldots, n$ where $n = 4{,}517$. We define $y_t = (y_{1t}, \ldots, y_{nt})'$. Five structural variables are available for each sales point, $x_{1i}, \ldots, x_{5i}$. They are time invariant and they have 2, 3, 4, 2 and 3 levels respectively. For each time period t the sales for nine products are available. The nine product sales are defined as $s_{jt}^{(i)}$ where $j = 1, \ldots, 9$ is the product index, t is the time period and i is the unit. For simplicity, we refer to the product j at time t for all units as the vector $s_{jt} = (s_{jt}^{(1)}, \ldots, s_{jt}^{(n)})$. Finally, only some of the sales point have received a certain marketing campaign during the first time period. The marketing campaign indicator $z_i = 0$ if the sales point i is in the control group, and one otherwise. We will use the terms marketing campaign and treatment interchangeably. From now on, to simplify the notation, we will use x_k, for $k = 1, \ldots, 5$ to represent the 4,517-dimensional vector of the values of the kth structural variable. We will use an analogous notation for z.

Initially we performed an exploratory analysis of the data. We observed that there are strong patterns due to all the covariates available. In particular, we noted an association between the sales of two different products during the first time period. A strong correlation was also apparent for the outcome and the product sales at sequential time periods. Moreover, we noted the highly skewed distribution of the product sales s_{it} for the sales points that do not sell all of the nine products in a time period. Finally, we tested for association between structural variables. The models that we propose in this paper were based on the extensive exploratory analysis that we carried out. See Akacha et al. (2009) for more details on the exploratory analysis.

In order to improve the spread of the data, the shrinkage of s_{j1} towards zero, due to the large range of its tails, was reduced by using the log transformation on the product sales s_{jt}. However, an issue arises when we apply the log transformation: there are several products with zero sales observed.

One of the main issues with the dataset is due to the design of the marketing campaign: for the sales points in the control group not all the possible configurations of the structural variables x_i have been observed. In particular, there are 61 (out of 144) combinations of the categories of the structural variables x for which we have no information for $z = 0$. This accounts for more than a third of the possible configurations and it will affect our results by restricting our ability to test for the effect of some factors on the impact of the marketing campaign.

Moreover, we note that the product sales are highly correlated. This does not allow us to include these covariates directly in the model matrix as we need it to be full rank. Thus we choose to use Principal Component Analysis (PCA). In particular, here we implement the 'projection pursuit' method developed by Croux et al. (2007) and based on robust estimation of covariance matrices. This method is especially useful for high dimensional data as ours. We apply PCA to the product sales s_{jt} and identify the principal components c_{kt} for $k = 1, \ldots, 9$.

The exploratory analysis of the data provided by the organisers uncovered the presence of a structure between the covariates and the outcome variable, and this structure provides us with the empirical motivation for the assumptions of the models proposed in Sects. 3 and 4. However, it also uncovered issues that require careful consideration in the modeling stage, such as an unbalanced design and a strong association between some of the covariates.

3 The Impact of the Marketing Campaign on Outcome at Time Period 1

The first question asked by the organisers is to evaluate the impact of the marketing campaign z on the outcome y_1. As this question involves only the data for the first time period, in this Section we drop the subscript t and use the notation $s_1,\ldots,s_9$ and y instead of $s_{11},\ldots,s_{91}$ and y_1.

Regression models provide us with tools to identify and estimate the impact of treatment. Although we could use a regression model that includes only the treatment as the only covariate, the availability of many other covariates, as usual in a data mining problem such as this one, allows us to study the impact of the marketing campaign once the confounding effect of the other covariates has been removed. Our exploratory analysis has shown that the structural variables are potential covariates of interest and interaction terms will also need to be included in our model fit. However, the existence of empty cells in the design of the experiment on the marketing campaign (see Sect. 2) restricts the inclusion of all the possible interactions.

The product sales are potential covariates of interest as well. However, there is a high frequency of units with product sales equal to zero, as discussed in Sect. 2. This motivates our proposal of a mixture model, given by

$$y_i = \xi g_{1i} + (1-\xi) g_{2i} \qquad \text{with } \xi \in [0,1], \tag{1}$$

where

$$g_{1i} = \mu_{1i} + \epsilon_1 \text{ with } \epsilon_1 \sim \mathcal{N}(0,\sigma_1^2), \text{ if } s_j^{(i)} > 0, \forall\, j = 1,\ldots,9, \tag{2}$$

or 0 otherwise, and

$$g_{2i} = \mu_{2i} + \epsilon_2 \text{ with } \epsilon_2 \sim \mathcal{N}(0,\sigma_2^2), \text{ if } \exists\, s_j^{(i)} = 0, j = 1,\ldots,9, \tag{3}$$

or 0 otherwise.

The MLE estimate of the proportion of products that are not sold during the first time period is given by $\hat{\xi} = 0.1348$ (with $SE(\hat{\xi}) = 0.0051$), a proportion of the data that cannot be ignored. Therefore, we propose different models for g_1 and g_2, to which we will refer as *group 1* and *group 2* respectively from now on. Further

analysis, not included here, also confirmed a different variance between the two groups, justifying $\sigma_1^2 \neq \sigma_2^2$.

The unbalanced design of the marketing campaign imposes restrictions on the interaction terms that can be included in the model for group 2. We performed stepwise model selection by AIC (Hastie and Pregibon 1992) to provide a measure of how well future outcomes are likely to be predicted by the model and to fit models to exclude non significant variables and interactions. This yields the final model which includes: z, $x_1, \ldots, x_5$, the interaction term (x_1, x_4), seven indicator functions for the product sales s_j with $j = 1, 2, 5, 6, 7, 8, 9$ (the indicator is equal to 1 when products of that category have been sold in the first time period), $\log(s_3)$, $\log(s_4)$ and two indicator functions for $s_5 > 75$ and for $s_7 > 75$. Note that all the sales points sold a positive amount of products j for $j = 3, 4$. Also, note that product sales s_5 and s_7 have a highly skewed distribution for the observations in group 2, motivating the use of an additional indicator function to differentiate the tails from the main body of their distribution. Therefore, the regression equation fitted for sales point i is given by

$$\begin{aligned}\mathbb{E}(g_{2i}) = {} & \alpha_0 + \alpha_z I(z_i = 1) + \alpha_1(2) I(x_{1i} = 2) + \ldots + \alpha_5(3) I(x_{5i} = 3) \\ & + \alpha_{1,4}(2,2) I(x_{1i} = 2, x_{4i} = 2) \\ & + \beta_1 I(s_1^{(i)} = 0) + \beta_2 I(s_2^{(i)} = 0) + \beta_5 I(s_5^{(i)} = 0) + \ldots + \beta_9 I(s_9^{(i)} = 0) \\ & + \beta_3 \log(s_3^{(i)}) + \beta_4 \log(s_4^{(i)}) + \gamma_5 I(s_5^{(i)} > 75) + \gamma_7 I(s_7^{(i)} > 75)\end{aligned}$$

where $I(.)$ is the indicator function, $\alpha_k(j)$ is the parameter corresponding to $I(x_{ki} = j)$ and $\alpha_{k,l}(j,h)$ is the parameter corresponding to $I(x_{ki} = j, x_{li} = h)$. The resulting model has 23 significant coefficients with an overall treatment effect of 32.74, while the estimates of the other regression parameters, together with their standard errors and p-values, are given in Akacha et al. (2009).

The exploratory analysis performed in Sect. 2 and the residual analysis shown in (Akacha et al. 2009) support our proposed model for group 2 that states that, for the sales points that do not sell all the products, the marketing campaign has an average effect of increasing the outcome by €32.74 million.

The covariates s_j for $j = 1, \ldots, 9$ are all strictly positive for group 1. However, as discussed in Sect. 2, the presence of a strong dependence between them encourages the use of PCA to extract from the product sales s_j a subset of orthogonal continuous covariates c_j. Based on the standard deviation decrease per number of principal component included in the model, it is apparent that at least the first six principal components should be included in the model, and we find that the first eight principal components give us the best fit using AIC to exclude non significant variables and interactions.

Therefore, our chosen model includes: z, x_1, x_2, x_3, x_4, x_5, several interaction terms and the principal components c_k for $k = 1, \ldots, 8$. For sales point i the regression equation is given by

$$\begin{aligned}\mathbb{E}(g_{1i}) = \alpha_0 &+ \alpha_z I(z_i = 1) + \alpha_1 I(x_{1i} = 1) + \ldots + \alpha_5(3) I(x_{5i} = 3) \\ &+ \alpha_{1,2}(2,2) I(x_{1i} = 2, x_{2i} = 2) + \ldots + \alpha_{z,5}(1,3) I(z_i = 1, x_{5i} = 3) \\ &+ \alpha_{1,2,3}(2,2,2) I(x_{1i} = 2, x_{2i} = 2, x_{3i} = 2) + \ldots \\ &+ \alpha_{2,3,4,5}(2,3,1,2) I(x_{2i} = 2, x_{3i} = 3, x_{4i} = 1, x_{5i} = 2) + \gamma_1 c_1^{(i)} \\ &+ \ldots + \gamma_8 c_8^{(i)}\end{aligned}$$

where $I(.)$ is the indicator function, $\alpha_k(j)$ is the parameter corresponding to $I(x_{ki} = j)$ and $\alpha_{k,l}(j,h)$ is the parameter corresponding to $I(x_{ki} = j, x_{li} = h)$ and so on. The estimates of the other regression parameters, together with their standard errors and p-values, are given in full in Akacha et al. (2009) along with the residual analysis.

The impact of the treatment campaign is significant with an average effect of increasing the outcome by around €32.4 million with the presence of the significant interaction term for (z, x_3), causing an increase on the impact of the marketing campaign when this is combined with certain values of the structural variable x_3.

4 Forecasting the Outcome for the Seventh Time Period

The second question asked by the organisers of the competition is to forecast the economic return for the seventh time period y_7. We believe that the most natural approach for forecasting in time series is based on the Bayesian paradigm, as in this approach the inference is updated as data becomes available over time (Pole et al. 1994). Statements about the uncertain future are formulated as probabilities conditioned on the available information. However, the first step in a forecasting problem is the construction of a suitable model based on analysis of the known development of the time series. Therefore, we propose here an extension to the regression model we introduced in question one in order to include a time evolution that allows us to forecast.

The organisers of the competition did not specify whether the time periods have constant length. We assume here that the time periods have constant length, do not overlap and are strictly sequential.

We propose two models and we use validation tools to choose the most appropriate model for forecasting in our context. We choose to hold out the last available time period, the sixth, for validation. The data which are not held out are used to estimate the parameters of the model, the model is then tested on data in the validation period, and finally forecasts are generated beyond the end of the estimation and validation periods. The outcome y has been observed for the sixth time period, so we can compare the predicted data with the observed data. We will then use the best model selected with the method above to forecast the outcome for the seventh time period. Of course, when we finally forecast the outcome for the seventh time period we use all the available data for estimation, that is, we also use the data available for the sixth time period.

We compare observed data with the predictions for different models by measuring the uncertainty using the Mean Squared Error for the predictions (MSE), the Mean Range of the 95% interval for the predictions (MR), and the Mean Interval Score (MIS) (Gneiting and Raftery 2007).

The first model we consider is a first-order autoregressive model (Chatfield 2003), usually referred to as AR(1), where the current outcome y_t depends on the previous outcome y_{t-1} for $t = 2, 3, 4, 5$. This model is motivated by the strong correlation between adjacent outcomes. Moreover, we propose to include in the AR(1) model the marketing campaign z for the first time period, as we have shown in the previous Section its strong impact on outcome for the first time period, and different within-sales point variances for the first time period and the remaining times intervals.

The second model we propose is a dynamic linear mixed model (Pole et al. 1994) which combines sales point-specific random effects with explanatory variables whose coefficients may vary over time. A dynamic model allows the inclusion of time-dependent parameters, a realistic assumption in our context. A Bayesian approach was used for both models proposed.

It appears that only the effect of x_1 on y_t varies as time passes, therefore we choose the corresponding parameter of x_1 to vary over time while the parameters of the remaining structural variables remain constant. We also decided to include autoregressive sales point-specific random effects α_{it} in order to quantify the variation between different sales points. This is an important component of the model as it aims to capture the effect of unit specific variables that we did not include in the model.

In summary, we propose the following model:

$$\begin{aligned} y_{it} = {} & \beta_{0t} + \beta_{1t} z_i + \beta_{2t} I(s_{2t}^{(i)} > 75) + \beta_{3t} I(s_{3t}^{(i)} > 75) + \beta_{4t} I(s_{5t}^{(i)} > 75) \\ & + \beta_{5t} I(s_{6t}^{(i)} > 75) + \delta_{1t} I(x_{1i} = 2) + \delta X_i + \alpha_{it} + v_t, \end{aligned} \tag{4}$$

where

$$\begin{aligned} v_t|\sigma_t^2 &\sim \mathcal{N}(0, \sigma_t^2), & \delta_{1t}|\delta_{1,t-1} &\sim \mathcal{N}(\delta_{1,t-1}, w_{\delta_1}^2) \\ \alpha_{it}|\alpha_{i,t-1}, w^2 &\sim \mathcal{N}(\alpha_{i,t-1}, w^2), & \beta_{kt}|\beta_{k,t-1} &\sim \mathcal{N}(\beta_{k,t-1}, w_k^2) \text{ for } k = \{0, 1, \ldots, 5\}, \end{aligned}$$

with $\sigma_1^2 \neq \sigma_2^2 = \ldots = \sigma_6^2 = \sigma^2$, $\delta = (\delta_2, \ldots, \delta_9)'$ and $X_i = I(x_{2i} = 2), I(x_{2i} = 3), I(x_{3i} = 2), I(x_{3i} = 3), I(x_{3i} = 4), I(x_{4i} = 2), I(x_{5i} = 2), I(x_{5i} = 3)'$ for $t = 2, \ldots, 5$ and $i = 1, \ldots, 4{,}517$. Furthermore, $w_{\delta_1} = w_k = 1$ for $k = \{0, 1, \ldots, 5\}$. The prior distributions chosen are reasonably non informative and are given in detail in Akacha et al. (2009) along with the variance component of the model. Samples may be generated from the model described using a Markov chain Monte Carlo (MCMC).

The assumption of different within-sales point variances in model 2 is confirmed by the model fitting. Moreover, the results show that the dynamics in the coefficients are very important for some of the parameters. We observe there that the overall

mean is monotonically increasing over time, whereas the treatment effect decreases as time passes. Also, it appears that the parameters associated with the indicators of the tails for the product sales s_{2t}, s_{3t}, s_{5t} and s_{6t} vary substantially over time. The time varying effect of x_i on the outcome is also confirmed by our results. Furthermore, the coefficients that do not vary over time ($\delta_k; k = 2, \ldots, 9$) are significant except for δ_9, corresponding to the effect of $I(x_{5i} = 3)$.

4.1 Validation Results and Predictions

We now compare observed data with the predictions obtained for the sixth time period from the two models proposed by measuring their uncertainty and precision. The validation result (omitted here, see Akacha et al. 2009) is that the dynamic model is superior to the autoregressive model using any of the three criteria.

From this validation analysis we conclude that the dynamic model gives better predictions and represents well the variability of the data. Thus, we proceed to predict data for the seventh time period using all the available data provided by the organisers.

We fit the observation (4) to predict the economic return during the seventh time period using all the data provided by the organisers. Drawing from the above distributions yields the predictions summarized in Fig. 1. See Akacha et al. (2009) for plots of the some of the predicted values and note how the dynamic model proposed captures the variability and the different shapes of the time series.

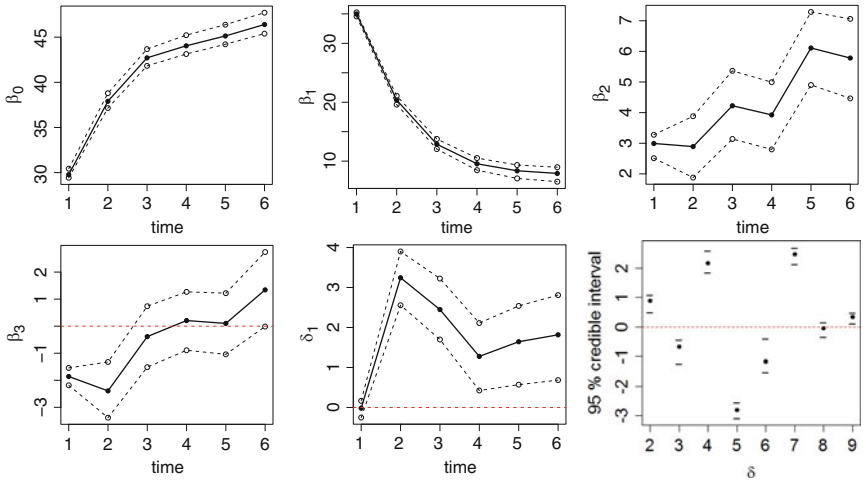


Fig. 1 95% credible intervals (*dashed line*) and median (*solid line*) for the mean parameters of model (4) using all the data available $y_1, \ldots, y_6$

5 Conclusion

In this paper we have analysed the data provided by the organisers of the competition and proposed an approach to answer the two questions. For the first question, we proposed a mixture of regression models while for the second question we performed this task within the Bayesian paradigm and proposed a dynamic model that, we believe, incorporates the main features of the data provided but it is also easy to adapt to the arrival of new information in real time, by updating the prior distributions or by including new covariates.

Acknowledgements We thank the SAS Institute for financial support and the organisers of the first Cladag Data Mining Prize for the constructive comments.

References

Akacha, M., Fonseca, T. C. O., & Liverani, S. (2009). *First CLADAG data mining prize: Data mining for longitudinal data with different marketing campaigns*. CRiSM Working Paper 09–46, University of Warwick. Available at http://www2.warwick.ac.uk/fac/sci/statistics/crism/research/2009

Chambers, J. M. (1992). Linear models. In J. M. Chambers & T. J. Hastie (Eds.), *Chapter 4 of statistical models in S*. Wadsworth & Brooks/Cole, Pacific Grove, California.

Chatfield, C. (2003). *The analysis of time series: An introduction*. CRC Pr I Llc, New York.

Croux, C., Filzmoser, P., & Oliveira, M. (2007). Algorithms for projection-pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory*, *87*, 218–225.

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*. Chapman & Hall/CRC, New York.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*, *102*(477), 360–378.

Hastie, T. J., & Pregibon, D. (1992). Generalized linear models. In J. M. Chambers & T. J. Hastie (Eds.), *Chapter 6 of statistical models in S*. Wadsworth & Brooks/Cole, Pacific Grove, California.

Pole, A., West, M., & Harrison, J. (1994). *Applied Bayesian forecasting and times series analysis*. Chapman & Hall/CRC, New York.

Part X
Data Analysis in Environmental and Medical Sciences

Supervised Classification of Thermal High-Resolution IR Images for the Diagnosis of Raynaud's Phenomenon

Graziano Aretusi, Lara Fontanella, Luigi Ippoliti, and Arcangelo Merla

Abstract This paper proposes a supervised classification approach for the differential diagnosis of Raynaud's Phenomenon on the basis of functional infrared imaging (IR) data. The segmentation and registration of IR images are briefly discussed and two texture analysis techniques are introduced in a spatial framework to deal with the feature extraction problem. The classification of data from healthy subjects and from patients suffering for primary and secondary Raynaud's Phenomenon is performed by using Stepwise Linear Discriminant Analysis (LDA) on a large number of features extracted from the images. The results of the proposed methodology are shown and discussed for images related to 44 subjects.

1 Introduction

Raynaud's Phenomenon (RP) is a paroxysmal vasospastic disorder of small arteries, pre-capillary arteries and cutaneous arteriovenous shunts of extremities, typically induced by cold exposure and emotional stress (Belch 2005). RP usually involves the fingers of the upper and lower extremities, even though tongue, nose, ears, and nipples may result affected as well. RP can be classified as primary (PRP), with no identifiable underlying pathological disorder, and secondary, usually associated with a connective tissue disease, the use of certain drugs, or the exposition to toxic agents (Block and Sequeira 2001). Secondary RP is frequently associated with systemic sclerosis. In this case, RP typically may precede the onset of other symptoms and signs of disease by several years (Belch 2005). It has been estimated that 12.6% of patients suffering from primary RP develops a secondary disease. In particular, while between 5 and 20% of subjects suffering from secondary RP evolves in either limited or diffuse systemic sclerosis, all of the systemic sclerosis patients underwent or will experience RP (Belch 2005). These epidemiological data point out the importance for early and proper differential diagnosis to distinguish the different forms of RP.

Thermal infrared (IR) imaging has been widely used in medicine to evaluate cutaneous temperature. IR imaging is a non-invasive technique providing the map of the

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_47,

superficial temperature of a given body by measuring the emitted infrared energy (Semmlow 2004). Since the cutaneous temperature depends on the local blood perfusion and thermal tissue properties, IR imaging provides important indirect information on circulation, thermal properties and thermoregulatory functionality of the cutaneous tissue. In this paper we thus exploit data from functional infrared imaging (fIRI) for classifying healthy controls (HCS), primary (PRP) and secondary (SSc) to systemic sclerosis RP patients.

The segmentation and registration of IR images are briefly discussed and texture analysis techniques are introduced in a spatial framework to deal with the feature extraction problem. Registration represents a crucial step of the analysis since most of the images are not aligned to each other. The classification is performed by using Stepwise Linear Discriminant Analysis (LDA) and the results of the proposed methodology are shown for a data set of 44 subjects. The paper is organized as follows. In Sect. 2 we describe the data and the processing of IR images giving specific details on the problems of image segmentation and registration. Sect. 3 considers the problem of feature extraction and describes two different procedures for performing texture analysis in a spatial framework: one based on the estimation of a Gaussian Markov Random Field and the other based on the calculation of texture measures obtained by co-occurrence matrices. Finally, in Sect. 4 we discuss classification results and in Sect. 5 we provide some conclusions.

2 The Processing of Thermal High Resolution Infrared Images

Data for this study were provided by the Functional Infrared Imaging Lab - ITAB, Institute for Advanced Biomedical Technology, at the School of Medicine of the G. d'Annunzio University, Chieti, Italy. For each subject, raw data consist of a temporal sequence of 46 images, each of dimension (256×256), documenting the thermal recovery from a standardized cold stress produced to the hands of each subject (Merla et al. 2002a). Specifically, we have $n = 44$ subjects classified as follows: 13 HCS, 14 PRP, and 17 SSc. The classification was performed according to the American College of Rheumatology criteria and standard exclusion criteria were observed (Merla et al. 2002a; Merla et al. 2002b).

The thermal high-resolution IR images were acquired every 30 seconds to monitor the response to a cold stress. To estimate the basal temperature of each subject, the image acquisition started 2.5 min before the cold stress and ended 20 min after. The cold stress consisted in a two minutes immersion of the hands in cold water (at 10°C), while wearing thin plastic gloves.

A discussion of classification results which exploits the information provided by the dynamic of the re-warming process can be found in Aretusi et al. (2010); here, we are interested in modelling the information provided by the first of the 46 images. In fact, by avoiding the cold stress, we might reduce not only time but also the financial cost of recording additional images.

It's often a necessary step, before a desired quantitative analysis, to carry out a processing of the images. In particular, in our experiment, prior to the feature

extraction for the classification of RP patients, we have to face with the problems of segmentation and registration of IR images. This is a crucial point of the analysis since for many subjects the images are not aligned at all.

Furthermore, since the nature of thermal images is quite different from that of a conventional intensity image,[1] it may happen that conventional segmentation algorithms may not be feasible when they are applied to a thermal images (Chang et al. 1997; Heriansyak and Abu-Bakar 2009).

The purpose of thermal image segmentation is to separate objects of interest from the background, usually represented by thermal features showing a certain degree of spatial uniformity. In such cases, it would be possible to perform a segmentation by using a threshold procedure. However, due to the slight blurring caused by the infrared imaging process, it usually happens that the boundary between a hand and the background is not so sharp for the images. In such cases, the images have to be segmented manually, pixel by pixel.

Due to the radically different ways of image formation in visible spectrum and in thermographic images, many methods for registration of images also work poorly or do not work at all (Zitova and Flusser 2003). The image registration is the process of geometrical alignment of two images, a sensed image with respect to a reference image, required to obtain more complete and comparable information through the subjects. For a survey on registration methods see, for example, Jarc et al. (2007). A reasonable way to practice, is first to manually detect the control points, usually by using an aided procedure. Then, a set of mapping function parameters, valid for the entire image, are estimated to align the reference and the sensed images. In general, a similarity transform, or an affine transform, is used in the mapping model; however, since in our study the distance, and the angle between the thermal camera and the scene are not always the same for all the subjects, a perspective projection model (Zitova and Flusser 2003), together with a bilinear interpolation method Pratt (2007), must be used to perform the image registration.

In Fig. 1 we show an example of image processing using 11 landmarks (control points) for the registration of the left hand segmented image.

Finally, in order to make the images spatially homogeneous, a reflection of the left hand with respect to its own longitudinal central axis is also needed.

3 Feature Extraction

With the aim of developing automatic discrimination techniques for HCS, PRP and SSc people, we have to extract a set of features from the registered images. Such images display complex patterns at various orientations and we thus expect quite

[1] In general, the latter encodes several physical properties such as reflectance, illumination and material of an object surface, to form the shape-related data, while a thermal image is formed by the heat distribution of an object; specifically, thermographic images depict the electromagnetic radiation of an object in the infrared range which is about 6–15 μm.

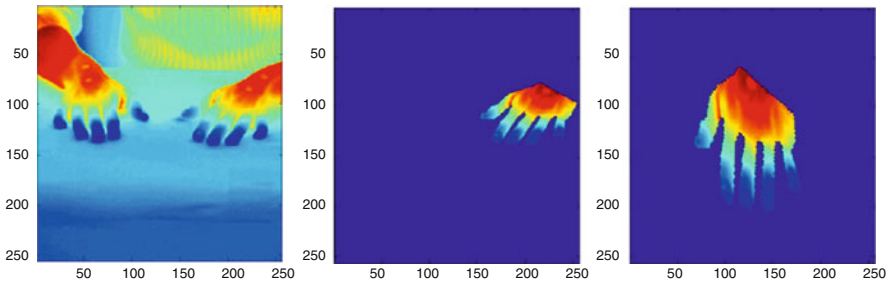


Fig. 1 A typical example of IR image: original image (*left*), segmented left-hand image (*middle*), registered image (*right*)

distinct texture characteristics among the classes. Texture analysis can be done either studying the point properties of an image, in a pixel-based view, or explicitly defining the primitives that characterize the image, in a structural approach, to search their features such as spatial arrangement. Thus, recalling that the relevant features for the supervised classification of the RP are unknown *a priori*, in this section we describe in detail the procedure used to generate our discriminant variables. Firstly, the temperature values are considered as a realization of a Gaussian Markov Random Field (GMRF) such that it is possible to assess the large and small scale variability of the process; the parameters of the GMRF constitute a subset of the vector of variables to be used in the classification process. Successively, further features of interest are defined from the co-occurrence matrices (CMs) computed on the residuals of the GMRF.

GAUSSIAN MARKOV RANDOM FIELDS. Consider an $N \times M$ image and let $X_{i,j}$ be the temperature level at pixel (i, j), $(i = 1, \ldots, N; j = 1, \ldots, M)$. We write $\mathbf{X} = (X_{1,1}, X_{1,2}, \ldots, X_{N,M})^T$ and take $n = NM$. The vector $\mathbf{X}$ contains the pixels in raster scan order-stacking the top row, then the second row, etc. The temperature values of the hands are considered as a realization of a Gaussian process characterized by a $(n \times 1)$ mean vector, $\boldsymbol{\mu}$, and a $(n \times n)$ covariance matrix, Σ. The mean structure can be modeled through a linear combination of independent variables with unknown parameters $\mathbf{b}$; i.e., $\boldsymbol{\mu} = \mathbf{Db}$. Specifically, if we consider a spatial trend-surface analysis, the entries of the design matrix $\mathbf{D}$ are expressed as a function of the coordinates of the pixels. Dealing with huge data sets, as in our case, it may be better for computational purposes to assume a conditional specification of the process. Accordingly, the temperature levels are distributed as a homogeneous Gaussian Markov Random Field if the distribution of X is multivariate normal with conditional means and variances

$$E(x_{i,j}|x_{r,s} : (r, s) \in \delta_{i,j}) = \mu_{i,j} + \sum_r \sum_s \beta_{r,s}(x_{i-r,j-s} - \mu_{r,s}),$$

$$\mathrm{Var}(x_{i,j}|x_{r,s} : r, s \in \delta_{i,j}) = v^2,$$

where $\delta_{i,j}$ is the set of neighbors of pixel (i, j), not including (i, j); $\beta_{r,s} = \beta_{-r,-s}$ and $\beta_{0,0} = 0$ are the spatial interaction parameters. For a first order stationary GMRF we have two spatial interaction parameters $\beta_{0,1}$, $\beta_{1,0}$ for neighbors which are one pixel apart vertically and horizontally respectively. For a second order stationary GMRF we have four spatial interaction parameters $\beta_{0,1}$, $\beta_{1,0}$ together with $\beta_{1,1}, \beta_{1,-1}$ for diagonally adjacent neighbors in North-East and South-East directions. This description is known as the conditional specification of a GMRF (Cressie, 1993) and we say that $\mathbf{X} \sim N(\mathbf{Db}, \Sigma)$, where $\Sigma = v^2\mathbf{A}^{-1}$, where $\mathbf{A}(\beta)$ is the $n \times n$ *potential matrix*, with entries equal to 1 along the principal diagonal, $-\beta_{(i,j),(r,s)}$ if the pixels (i, j) and (r, s) are neighbors, and zero otherwise. Note that $\mathbf{A}(\beta)$ must be strictly positive definite for a non-degenerate distribution. The primary objective is to estimate the parameters of the q-vector β containing the distinct $\beta_{r,s}$ parameters, τ^2 (the conditional variance) and the vector of trend parameters, $\mathbf{b}$. An important part of the GMRF model specification is the choice of boundary conditions (b.c.) for a stationary process, since elements of Σ^{-1} for boundary sites on a finite lattice can be very complicated (Besag and Moran 1975). In general, to deal with GMRFs on finite rectangular lattices, the most convenient boundary conditions are toroidal b.c. These specify that each dimension is assumed to be wrapped around, so that the first and last coordinates are adjacent. There are many possible methods for estimating the parameter vector $\boldsymbol{\eta} = (\mathbf{b}, \beta, v^2)$. If toroidal boundary conditions are assumed then $\mathbf{A}(\beta)$ is block circulant and can be easily diagonalized with a two dimensional fast Fourier transform (Besag and Moran 1975). If $\mathbf{P}_j$ is the usual discrete Fourier transform matrix, then $\mathbf{P} = \mathbf{P}_N \otimes \mathbf{P}_M$ is the $n \times n$ matrix of the eigenvectors of $\mathbf{A}(\beta)$, with $\otimes$ being the Kronecker product. The eigenvalues of $\mathbf{A}(\beta)$ are simple to obtain from the 2D discrete Fourier transform and we write the eigenvalues as $q_1, \ldots, q_n$. Hence, $\mathbf{A}(\beta) = \mathbf{PQ}(\beta)\mathbf{P}^T$, where $\mathbf{Q}(\beta) = \text{diag}(q_1, \ldots, q_n)^T$, and $\mathbf{A}(\beta)$ is positive definite if all $q_i > 0$. The log-likelihood is

$$L(\mathbf{b}, \beta, v^2|\mathbf{x}) = -\frac{n}{2} log(2\pi\, v^2) + \frac{1}{2}\sum_{i=1}^{n} \log(q_i) - \frac{1}{2v^2}(\mathbf{x}-\mathbf{Db})^T\mathbf{PQ}(\beta)\mathbf{P}^T(\mathbf{x}-\mathbf{Db})$$

and the maximization can be carried out with only $O(n \log n)$ steps (Besag and Moran 1975; Dryden et al. 2002) over the valid parameter space, i.e., subject to $\mathbf{A}(\beta)$ being strictly positive definite.

Texture Analysis. Here, we use a pixel-based approach to identify further basic patterns that could represent the natural texture structure of the RP. Specifically, we perform a texture analysis by extracting information in the form of a co-occurrence matrix (CM) and by summarizing this information through the calculation of some measures of texture on the CM (Cocquerez and Philipp 1995). To calculate these measures, for each image we first classify the estimated stationary residual process, $\hat{\varepsilon}_{ij} = x_{ij} - \mu_{ij}$, in L levels, where L is chosen by considering the quantiles of the distribution. Then, for a given spatial displacement vector, $\mathbf{r}$, which defines pairs of neighbors in the spatial domain, we compute the CM which provides a tabulation of how often different combinations of classified pixel values occur in an image

(Cocquerez and Philipp 1995). More specifically, the $(i, j)th$ element of the $(L \times L)$ CM, denoted here as $\mathbf{C_r}$, represents the estimated probability, $f(i, j)$, of occurrence of a pair of classified pixel values, separated by the displacement $\mathbf{r}$ and having temperature levels i and j, respectively. Therefore, for a displacement vector $\mathbf{r}$, we calculate the set of the following texture measures

$$T_1(\mathbf{r}) = \sum_{i,j} (i-j)^2 f(i,j), \quad T_2(\mathbf{r}) = \sum_{i,j} \frac{f(i,j)}{1+|i-j|},$$

$$T_3(\mathbf{r}) = \sum_{i,j} f(i,j)^2, \quad T_4(\mathbf{r}) = \frac{\sum_{i,j} i\ jf(i,j) - \sum_i if(i,\cdot) \sum_j jf(\cdot,j)}{\sigma_i\ \sigma_i},$$

$$T_5(\mathbf{r}) = \sum_{i,j} f(i,j) \log_2 \frac{f(i,j)}{f(i,\cdot) f(\cdot,j)}$$

where $\sigma_i = \left[\sum_i i^2 f(i,\cdot) - (\sum_i if(i,\cdot))^2\right]^{1/2}$, $\sigma_j = \left[\sum_j i^2 f(\cdot,j) - (\sum_j jf(\cdot,j))^2\right]^{1/2}$ and $f(\cdot, j)$ and $f(i,\cdot)$ represent the marginal probabilities over the indices j and i, respectively. The indices T_1 and T_2 represent *Contrast* and *Homogeneity* measures and use weights related to the distance from the diagonal of the CM; T_3 is known as *Energy* and gives information about orderliness; finally, T_4 and T_5 are *Correlation* and *Mutual Information* indices, respectively; they provide a measure of the linear and non linear dependence of pairs of classified pixel values.

4 Classification Results

In this section we discuss discrimination results on the data described in Sect. 2. For computational purposes, we perform the procedures of parameter estimation and feature extraction on left and right hand images. Hence, for each hand, we have 44 images, each of dimension (128×128). In total, we have $n = 44$ subjects classified as HCS, PRP, and SSc. The identification of the feature variables, for each subject, starts with the estimation of the parameters of a GMRF which is very commonly used for modeling textures in image analysis (Cressie 1993; Dryden et al. 2002). The estimated mean function is based on a 6-parameter spatial quadratic trend expressed as a polynomial function of spatial coordinates. For the residual correlated process, we thus consider a neighborhood structure with four neighbors in space (i.e., second order GMRF); this corresponds to a GMRF with 5 parameters (including the conditional variance) to be estimated. However, notice that this procedure leads to the estimation of 22 parameters for the whole image with both hands. As regards the use of co-occurrence matrices, we consider twenty levels ($L = 20$) and spatial displacements corresponding to the four main spatial directions (i.e., East, West, North, South). Considering a spatial lag up to 8, all the displacements

can be collected in a global vector, $\mathbf{d}$, which takes the following structure: $\mathbf{d} = [(0\ 1); (0\ 2); \ldots; (0\ 8); (-1\ 1); (-2\ 2); \ldots; (-8\ 8); (-1\ 0); (-2\ 0); \ldots; (-8\ 0); (-1\ -1); (-2\ -2); \ldots; (-8\ -8)]$. Thus, overall we specify 32 different spatial lags and, for each of them, we can calculate the corresponding CMs. However, to avoid an increase of the number of discriminant variables, for each spatial displacement, $\mathbf{r}$, and for each pair of levels (i, j), we aggregate results for both hands thus obtaining a synthesized CM matrix, $\tilde{\mathbf{C}}_{\mathbf{r}}$, from which we can calculate the five texture measures, $T_1, \ldots, T_5$. This procedure, generates 160 variables, and by joining them with the ones provided by the specification of the GMRF, for each subject we thus have a total of 182 variables.

Of course, it is highly likely that a large number of these features do not provide any significant discriminatory information. Hence, to reduce the number of variables to a suitable number for the classification routine we use a forward stepwise linear discriminant analysis (Johnson and Wichern 2007). The best subset of selected features varies according to the discrimination criterion. For example, by using the *smallest F ratio* (SFR) criterion a total of 13 discriminant variables are selected, while using the Mahalanobis distance (MD) the number of these variables increases up to 16. The selected variables are mainly related to the vertical and diagonal directions, with only a few horizontal lags chosen. In general, two or three of the selected variables are related to the interaction parameters of the GMRF while the remaining ones are represented by the indices $T_1 - T_5$. Variables related to the diagonal directions are characterized by a displacement of 2–4 lags while variables related to the vertical one are defined for spatial lags ranging from 3 to 8. The wider spatial lags observed for the vertical direction could be likely linked to the specific geometry of the finger vasculature and the expression of the functional impairment secondary to the disease. In fact, larger finger vessels run longitudinally and parallelously, while only a few artero-venous shunts run transversally. Moreover, the presence of scleroderma leads to a progressive destruction of the microvasculature from distal to proximal sections, thus explaining the differences observed in the spatial lags. The p-values (≈ 0) for the Wilk's Lambda statistics show that the differencies in the group mean discriminant scores are greater than what could be attributed to sampling error. The mean vectors lie mainly in the first discriminant dimension (the proportion of the first eigenvalue on the total is 58,9% for SFR and 51,6% for MD). The canonical correlations between each of the two discriminant functions and the grouping (dummy) variables are very high both for SFR (0,908 and 0,875) and MD case (0,941 and 0,937). The confusion matrix resulting from LDA based on the *smallest F ratio* and the Mahalanobis distance criteria gives a 0% estimate of the apparent error rate, but performing the Leave-One-Out Cross Validation (LOCV) procedure, the estimated error rate increases up to 15.9% (SFR) and 13.6% (MD); this corresponds to a misclassification of 3 HCS, 2 PRP and 2 SSc for SFR and 1 HCS, 1 PRP and 4 SSc for MD. In general, according to additional clinical information, it results that in both cases the misclassified PRP are affected by other clinical forms (i.e., systemic lupus erythematosus and bilateral carpal trauma) and that one of the misclassified SSc is also affected by a connective tissue disease that may be characterized by particular functional aspects.

5 Conclusions

In this paper we have discussed the classification problem of Raynaud's Phenomenon on the basis of functional infrared imaging (IR) data. The classification procedure is complex and deals with problems in terms of segmentation, image registration and feature extraction. Classification results have been obtained through texture analysis. As shown in Aretusi et al. (2010), the classification can be improved by exploiting the information provided by the dynamic of the re-warming process: however, to our knowledge, there are no other works which achieve our results based on the analysis of one single image.

Despite the good results achieved, there are still some open problems to be considered; for example, the study of the performance of different classifiers (e.g., Discriminant Partial Least Square and Lasso Discriminant Analysis) as well as the comparison of different features selection algorithms may be of interest and this will be the object of future works.

References

Aretusi, G., Fontanella, L., Ippoliti, L., & Merla, A. (2010). *Space-Time texture analysis in thermal infrared imaging for classification of Raynaud's phenomenon. In Complex data modeling and computationally intensive statistical methods*. Contribution to Statistics Series, Springer, 1–12. Eds. Mantovan P. and Secchi P.

Belch, J. (2005). Raynaud's phenomenon. Its relevance to scleroderma. *Annals of the Rheumatic Diseases*, *50*, 839–845.

Besag, J. E., & Moran, A. P. (1975). On the estimation and testing of spatial interaction in Gaussian lattice processes. *Biometrika*, *62*, 555–562.

Block, J. A., & Sequeira, W. (2001). Raynaud's phenomenon. *Lancet*, *357*, 2042–2048.

Chang, J. S., Liao, H. Y. M., Hor, M. K., Hsieh, J. W., & Cgern, M. Y. (1997). New automatic multi-level thresholding technique for segmentation of thermal images. *Images and Vision Computing*, *15*, 23–34.

Cressie, N. A., (1993). *Statistics for spatial data* (2nd ed.). New York: Wiley.

Cocquerez, J. P., & Philipp, S. (1995). *Analyse d'images : Filtrage et segmentation*. Paris: Masson.

Dryden, I. L., Ippoliti, L., & Romagnoli, L. (2002). Adjusted maximum likelihood and pseudo-likelihood estimation for noisy Gaussian Markov random fields. *Journal of Computational and Graphical Statistics*, *11*, 370–388.

Heriansyak, R. & Abu-Bakar, S. A. R. (2009). Defect detection in thermal image for nondestructive evaluation of petrochemical equipments. *NDT&E International*, *42*, 729–740.

Jarc, A., Pers, J., Rogelj, P., Perse, M., & Kovacic, S. (2007). Texture features for affine registration of thermal and visible images. Computer Vision Winter Workshop.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Prentice Hall, London: Pearson education.

Merla, A., Romani, G. L., Di Luzio, S., Di Donato, L., Farina, G., Proietti, M., Pisarri, S., et al. (2002a). Raynaud's phenomenon: Infrared functional imaging applied to diagnosis and drug effect. *International Journal of Immunopathology and Pharmacology*, *15*(1), 41–52.

Merla, A., Di Donato, L., Pisarri, S., Proietti, M., Salsano, F., Romani, G. L. (2002b). Infrared functional imaging applied to Raynaud's phenomenon. *IEEE Engineering in Medicine and Biology Magazine*, *6*(73), 41–52.

Pratt W.K. (2007). Digital Image Processing. John Wiley and Sons, Hoboken, New Jersey.
Semmlow, J. L. (2004). *Biosignal and biomedical image processing*. London: CRC Press.
Zitova, B., & Flusser, J. (2003). Image registration methods: a survey. *Image and Vision Computing*, *21*, 977–1000.

A Mixture Regression Model for Resistin Levels Data

Gargano Romana and Alibrandi Angela

Abstract Resistin is a mainly adipose-derived peptide hormone, that reduces insulin sensitivity in adipocytes, skeletal muscles and hepatocytes. Only in recent years resistin has been studied in liver disease and the regarding literature is very poor. According to recent studies considering resistin like a clinical biomarker in the assessment of liver cirrhosis, we propose an application of a finite mixture regression model with concomitant variable in order to individualize factors that influence resistin levels in patients affected by different virus hepatitis. The estimated model shows the existence of two separated components differing for the intercept and for some covariates; moreover high serum resistin levels do not seem to be associated with liver histological lesions by C virus, but only by B virus hepatitis.

1 Introduction

Resistin is a recently discovered adipocytokine and it seems to play a role in glucose homeostasis, insulin resistance and inflammation (Tiftikci et al. 2009). It serves as a signaling molecule between energy storage organ, adipose tissue, and the principal insulin-responsive organs, liver, muscles and fat. In humans, the function of resistin is not yet fully elucidated, but it seems to have a role in the inflammatory response rather than in the regulation of glucose homeostasis. Resistin has been shown to influence insulin sensitivity in various tissues, but the physiological function in the liver disease is still controversial.

Only a few studies have investigated resistin relevance in pathophysiology of liver injury. Recent researches have indicated that resistin is expressed within human liver and its expression increases during severe damages, such as alcoholic liver disease and hepatitis. Resistin expression during liver damage is positively correlated with histological parameters of inflammation, suggesting that inflammatory cells may represent the principal cell type responsible for intrahepatic resistin production (Bertolani et al. 2006). Another study showed that resistin is elevated in states of critical illness, even without evident infection (Koch et al. 2009).

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_48,

The possible relevance of resistin in chronic liver disease has been confirmed by studies which analyzed plasma resistin concentrations in patients with liver cirrhosis. Serum resistin levels were elevated in patients with cirrhosis compared to healthy subjects and were dependent on the clinical stage of the disease, as well as on the presence of an inflammatory state (Yagmur et al. 2006). All these studies have modeled resistin essentially by means of ANOVA (Yang et al. 2009), non parametric tests, Cox regression and Kaplan- Meier estimate (Koch et al. 2009).

In this study the effects of hepatitis B virus (HBV) and hepatitis C virus (HCV) infection on plasma resistin levels were evaluated. This pathology represents one of the most serious public health problems and the main cause of chronic liver disease worldwide. A mixture regression model with concomitant variable was modelled in order to individualize the possible factors influencing resistin serum levels in patients affected by different type virus hepatitis. These models are suitable when the patients are not alike. In a therapeutic setting patients differently react to treatment regimes and this may depend on known or unknown factors. In the last case there is an unobserved patients heterogeneity since it is not possible to observe directly to which subpopulation a patient belongs. Likewise, the underlying covariate causing variability in treatment response is unknown.

This paper is organized as follows. Section 2 presents an outline of finite mixture regression models, the identifiability, the estimation and the most popular criteria for identifying and choosing components number. Section 3 shows the model application to the resistin data and Sect. 4 discusses some concluding remarks.

2 Finite Mixture Regression Models

Finite mixture regression models are methods to model unobserved heterogeneity or to account overdispersion in data. They have been extensively discussed in the literature and applied in various areas (McLachlan and Peel 2000).

Mixture models can be expressed as a weighted sum of K components, each component follows a parametric distribution and has an assigned weight indicating the a priori probability for an observation to come from this component. If the weights depend on further variables, these are referred to as concomitant variables (Grün and Leisch 2008).

A general model class is given by:

$$f(y|x,\omega,\Theta) = \sum_{i=1}^{K} \pi_i(\omega,\alpha) f_i(y|x,\theta_i) \tag{1}$$

where Θ denotes the vector of all parameters for the mixture density $f(\cdot)$, y denotes the response, x the predictors, ω the concomitant variable and α the parameters of the concomitant variable, $f(\cdot)$ is the component specific density function. The concomitant specific parameters are given by θ_i that in the Gaussian distribution are

given by $\theta_i = (\beta_i', \sigma_i^2)$ where β_i are the regression coefficients and σ_i^2 represents the variance. The mixing proportions or component weights $\pi_i(\omega;\alpha)$ are the probabilities in a multinomial distribution consisting of one draw on K categories. They obey $\sum_{i=1}^{K} \pi_i(\omega;\alpha) = 1$; $\pi_i(\omega;\alpha) \geq 0$, $\forall i$ and they are usually modelled as functions of the covariates by the logistic function.

In general the concomitant variable is assumed to be a multinomial logit model:

$$\pi_i(\omega,\alpha) = \frac{e^{\omega'\alpha_i}}{\sum_{s=1}^{K} e^{\omega'\alpha_i}} \quad \forall i, \tag{2}$$

with $\alpha = (\alpha_i')$ and $\alpha_1 \equiv 0$.

The parameter vectors of a statistical model are identifiable if two non-equivalent parameter vectors parameterize the same distribution. This definition depends on the notion of parameter vectors 'equivalence'. In mixture modelling, two parameter vectors are equivalent if they are equal up to permutation of the mixture components.

Generic identifiability is guaranteed for important continuous distributions such as the Gaussian. The identifiability of mixtures of Gaussian regression models is analyzed in Hennig (2000). The results indicate that requiring a covariate matrix of full rank is not sufficient. Contrarily, it is necessary to check a coverage condition in order to ensure identifiability. With respect to generic identifiability of finite mixtures of regression models, three influencing factors can be distinguished: component distribution f, covariate matrix and repeated observations/labelled observations. Hennig showed that mixtures of Gaussian linear regressions with assignment independence are identifiable if the number of 'lower than p'-dimensional hyperplanes needed to cover all the x is larger than the number of mixture components.

For mixtures where the component distributions are identifiable this means that the component weights and possible dispersion parameters are unique, but the regression coefficients vary because they depend on the combination of the components between the covariate points. This identifiability problem is also of concern for prediction, because given the class membership the predicted value for new data depends on the chosen solution.

There exist different methods for estimation of finite mixture models. The most popular is the EM algorithm (Dempster et al. 1977) which aims at determining the ML estimator for a finite mixture model with a given number of components K. The EM algorithm has the advantage of providing a general framework for estimating different kinds mixture models as often only the M-step has to be modified if different component specific models are used. In addition, already available tools for weighted maximum likelihood estimation can be applied.

An important characteristic of the estimation method is if the number of components has to be fixed a-priori or is simultaneously estimated. When the number of mixtures is unknown, we could estimate as follows. Use the EM algorithm to obtain a sequence of parameter estimates for a range of K values and estimate $\hat{K}$ as:

$$\hat{K} = \arg\min_{k} \mathscr{C}(\hat{\beta}(x), k) \; k = k_{min}, \ldots, k_{max} \tag{3}$$

where $\mathscr{C}(\cdot; k)$ is some model selection criterion. There are many choices for $\mathscr{C}(\cdot; k)$. Most of approaches use penalty term including Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC). This implies the introduction of penalty terms for the number of parameters estimated for the model at hand. Some approaches use the log-likelihood-ratio test to assess the hypothesis (H_0) that the sample is from K_0 components mixture distribution again the alternative hypothesis (H_1) that the sample is draw from a $K_0 + 1$ components mixture distribution. Another line of research tried to use the parametric bootstrap methods (McLachlan and Basford 1988 and McLachlan 1987) and procedure based on nonparametric bootstrap using a mixture algorithm which combines the Vertex Exchange and the EM-algorithm (Richardson and Green 1997). Moreover, there exist methods for the simultaneous determination of the number of components and variables in finite mixture regression models, as the Mixture Regression Criterion, that yields marked improvement in model selection due to clustering penalty function (Naik et al. 2007).

3 A Model for Resistin Levels Data

A total of 118 consecutive subjects aged 31–70 years diagnosed with hepatitis B or C virus infection and referred to Unit of Clinica Medica at Messina University (Italy) were enrolled between January 2007 and January 2009.

We examined 32 (24 males and 8 females) patients with HBV and 86 (48 males and 38 females) with HCV infection. Exclusion criteria were: (a) other causes of chronic liver disease, (b) decompensated liver disease, (c) alcohol intake > 20 g/day during the previous 6 months, (d) seropositivity for anti-HIV, (e) history of heart failure, diabetes, thyroid diseases, abnormal renal function and cancer, (f) morbid obesity diagnosed by means of body mass index (BMI) > 40).

All patients underwent a laboratory test including measurements of resistin (HBV: 5.20 ± 1.09 ng/mL, HCV: 2.51 ± 1.32 ng/mL), total cholesterol (HBV: 160 ± 53 mmol/L, HCV: 175 ± 43 mmol/L), triglycerides (HBV: 114 ± 51 mg/dl, HCV: 96 ± 42 mg/dl), insulin (HBV: 23 ± 22 ng/ml, HCV: 19 ± 12 ng/ml), glycemia (HBV: 81 ± 22 mg/dl, HCV: 94 ± 39 mg/dl), BMI (HBV: 24.01 ± 3.06, HCV: 25.86 ± 2.71). Insulin resistance was determined using the HOMA index (Homeostasis Model Assessment) (HBV: 5.20 ± 6.66, HCV: 4.83 ± 4.62).

In literature the adipokines (including resistin) levels have been analyzed by means of non parametric test (to asses the differences between groups) and multiple regression models (to evaluate the possible predictors).

Preliminary, according to literature, a regression model in which the dependent variable was serum resistin levels and independent variables were all the other considered variables was estimated. It showed that resistin concentration was significantly dependent only on triglycerides ($p = 0.016$) with a low R^2 value (0.412). This model did not take in account the belonging of each subject to a specific group or type of hepatitis.

Table 1 The estimated logLik, AIC, BIC and ICL for the resistin levels data

K	logLik	AIC	BIC	ICL
1	−245.786	515.572	550.165	550.253
2	−163.526	403.053	412.599	518.176
3	−187.855	425.710	497.779	490.021

In this study, the resistin serum by mixture regression model with concomitant variable was modelled in order to individualize the possible factors influencing resistin levels in patients affected by different type virus hepatitis. This choice is due to the awareness that the resistin levels may be different in patients with HBV and HCV and they may depend on different covariates, within each subgroup. The mixture regression models represent the methodologically adequate solution, because it allows us to study if the serum resistin levels depend on different factors in patients with different hepatitis; in particular, we evaluate the effect of hepatitis B virus and hepatitis C virus infection.

Table 1 reports the estimated log-likelihood (LogLik), AIC value, BIC value and Integrated Complete Likelihood (ICL) determined in order to identify the number of mixture components, each of them including all the regressors above.

Log-likelihood, AIC and BIC criteria suggest that the two-component gaussian mixture regression model provides a slightly better fit than others. It's important to note that the mixture models presented in the table possess a same set of regressors in each component.

The following step was to identify regressors by using the BIC again from the our two-component mixture regression model identified above.

We started with the most general model to determine the number of components using information criteria and a possible model restriction was checked after having the fixed components number.

Different mixture models with two components were fitted:

- a model with only intercept;
- a model with all covariates ignoring the group (B or C virus);
- a model with all covariates considering the group (B or C virus) as concomitant;
- different models with only few covariates and the group as concomitant.

The BIC value was used for models comparison. We preferred the BIC smaller model with concomitant variable including the intercept and five covariates: age, HOMA (that allows to reach a baseline estimate of insulin resistance), insulin (protein hormone produced by cells within the pancreas; it is secreted when the level of blood glucose is high), triglycerides (esterified with three fatty acids; high levels of triglycerides are associated with atherosclerosis and heart disease, risk of pancreatitis-inflammation and hepatitis), cholesterol (base substance for the synthesis of steroid hormones, it allows cell growth and division, cholesterol produced in the liver is largely used for the production of bile).

Table 2 shows the results of estimated model for each of two components.

Table 2 Summary of two components mixture regression model

Component 1				
	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	2.167	0.324	6.681	0.000
Cholesterol	−0.001	0.002	−0.737	0.461
Triglycerides	0.009	0.003	2.998	0.003
Insulin	−0.015	0.015	−0.990	0.322
HOMA	−0.030	0.031	−0.943	0.346
Component 2				
	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	5.054	0.277	18.233	0.000
Cholesterol	−0.015	0.002	−7.178	0.001
Triglycerides	0.018	0.003	6.019	0.003
Insulin	0.149	0.023	6.468	0.002
HOMA	−0.575	0.082	−6.993	0.000

It is important emphasize that the first subpopulation is mainly composed by patients affect by C virus infection and the second component mostly by subjects with B virus infection. The components differ for the intercept and some covariates. For the first component only triglycerides are significantly different from zero, while for the second component all covariates are significant. In particular, cholesterol and HOMA are negatively associated with high resistin levels and the coefficient referred to HOMA index is higher than the cholesterol one, denoting a greater influence exerted by this regressor. Triglycerides and insulin are positively associated to the dependent variable, with a strong influence of insulin.

Figure 1 shows four residual plots for a single mixture component at time with all points.

These plots confirm that the components are very well separated.

4 Final Remarks

In this paper we proposed an application of a mixture regression model with concomitant variable allowing us to show that the high serum resistin levels do not seem to be strongly associated with liver histological lesions by C virus, but prevalently by B virus hepatitis. Results showed indeed that the two components of our model are well separated, so that we can affirm that the resistin serum levels are significantly different according to the existence of B or C virus infection. Our mixture regression model showed that resistin serum is influenced by different factors in the two subpopulations of subjects. In the first component, that is essentially composed by C virus infected, only triglycerides have a significant and positive influence on resistin levels variation. In the second component, prevalently composed by B virus infected, all regressors inserted in the model are statistically significant: resistin

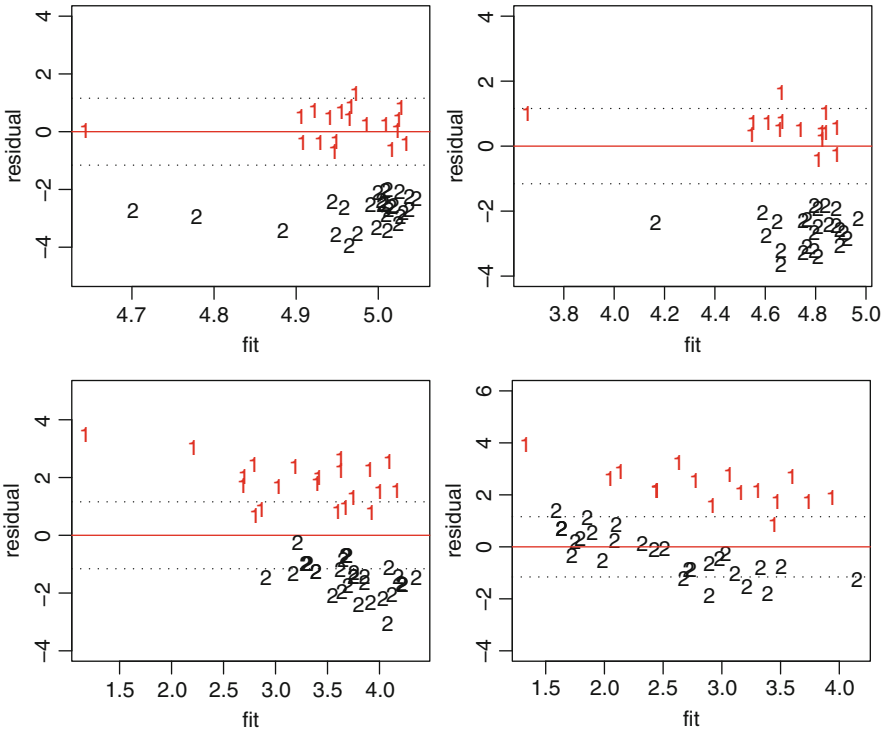


Fig. 1 Component-wise residual plots

serum positively depends on insulin and triglycerides, while negatively depends on cholesterol and, above all, HOMA index.

We have to emphasize that insulin and HOMA index are linked to diabetes and in medical literature is known the association between B virus hepatitis and diabetes.

Other studies are needed for defining better resistin role at the level of various organs and apparatuses and for confirming the link between resistin and B virus hepatitis. So, it might be possible to study drugs contrasting the effects of this hormone and early detection of diabetes development through the monitoring of resistin concentration.

In conclusion the use of mixture regression models for the analysis of heterogeneous populations, such as medical casuistry, can be useful to explain part of heterogeneity with known explanatory covariates. In fact in these cases the application of standard statistical models, such as linear regression, model the population average, which is the mean response of all individuals, without considering the existence of possible subpopulations. Moreover, when the usual assumptions associated with general linear models are suspect, the mixture of distributions model can be a possible alternative because they are less restrictive than the usual distributional assumptions and provide an alternative to nonparametric modeling.

References

Bertolani, C., Sancho-Bru, P., Failli, P., Bataller, R., Aleffi, S., DeFranco, R., Mazzinghi, B., Romagnani, P., Milani, S., Ginés, P., Colmenero, J., Parola, M., Gelmini, S., Tarquini, R., Laffi, G., Pinzani, M., & Marra, F. (2006). Resistin as an intrahepatic cytokine: Overexpression during chronic injury and induction of proinflammatory actions in hepatic stellate cells. *American Journal of Pathology*, *169*, 2042–2053.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from imcomplete data via the E-M algorithm. *Journal of the Royal Statistical Society B*, *39*, 1–38.

Grün, B., & Leisch, F. (2008). Finite mixtures of generalized linear regression models. In Shalabh and Christian Heumann (eds), *Recent advances in linear models and related areas* (pp. 205–230). Heidelberg: Physica-Verlag. *Computational Statistics and Data Analysis*, *51*, 5247–5252 (2007).

Hennig, C. (2000). Identifiability of models for clusterwise linear regression. *Journal of Classification*, *17*(2), 273–296.

Koch, A., Gressner, O. A., Sanson, E., Tacke, F., & Trautwein, C. (2009). Serum resistin levels in critically ill patients are associated with inflammation, organ dysfunction and metabolism and may predict survival of non-septic patients. *Critical Care*, *13*(3), R95.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics, 36*(3), 318–324.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*. New York: Marcel Dekker, Inc.

McLachlan, G., & Peel, D. (2000). *Finite mixture model*. New York: Wiley.

Naik, P. A., Shi, P., & Tsai C. L. (2007). Extending the akaike information criterion to mixture regression models. *Journal of the American Statistical Association*, *102*(477), 244–254.

Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society B*, *59*(4), 731–792.

Tiftikci, A., Atug, O., Yilmaz, Y., Eren, F., Ozdemi, F. T., Yapali, S., Ozdogan, O., Celikel, C. A., Imeryuz, N., & Tozun, N. (2009). Serum levels of adipokines in patients with chronic HCV infection: Relationship with steatosis and fibrosis. *Archives of Medical Research*, *40*, 294–298.

Yagmur, E., Trautwein, C., Gressner, A. M., & Tacke, F. (2006). Resistin serum levels are associated with insulin resistance, disease severity, clinical complications, and prognosis in patients with chronic liver diseases. *American Journal of Gastroenterology*, *101*, 1244–1252.

Yang, Y., Xiao, M., Mao, Y., Li, H., Zhao, S., Gu, Y., Wang, R., Yu, J., Zhang, X., Irwin, D. M., Niu, G., & Tan, H. (2009). Resistin and insulin resistance in hepatocytes: Resistin disturbs glycogen metabolism at the protein level. *Biomedicine and Pharmacotherapy*, *63*, 366–374.

Interpreting Air Quality Indices as Random Quantities

Francesca Bruno and Daniela Cocchi

Abstract Synthetic indices are a way of condensing complex situations to give one single value. A very common example of this in environmental studies is that of air quality indices; in their construction, statistics is helpful in summarizing multidimensional information. In this work, we are going to consider synthetic air-quality indices as random quantities, and investigate their main properties by comparing the confidence bands of their cumulative distribution functions.

1 Introduction

This study aims to identify genuinely different air quality situations by evaluating the uncertainty of air quality indices. As we pointed out in Bruno and Cocchi (2007), synthetic environmental indices explicitly include uncertainty. Recent studies of air quality indices have adopted diverse approaches, like the search for optimal indices as power averages in Ruggeri et al. (2009), or the Bayesian approach to estimating air quality indices able to include measures of variability, in Lee et al. (2009). Moreover, in Lagona and Maruotti (2009), air quality indices are used to study exceedances over thresholds, and to identify extreme pollution situations.

Air quality indices summarize pollution in a given area, and enable comparisons between areas to be made over the same timescale. In order to evaluate the differences between different areas, air quality indices ought to be considered random quantities in order to take account of their variability. In fact, they are functions of components that should be considered random, since they are outcomes of random processes and are affected by measurement errors, so their complex structure is better represented by a probabilistic model rather than a descriptive function. Information from synthetic indices can be enhanced by studying their distribution and by computing the probability of obtaining values that are higher or lower than fixed thresholds. When the focus is on an index that represents the less favorable pollution situations, the Generalized Extreme Value (GEV) distribution (Coles 2001, Beirlant et al. 2004) is a suitable probabilistic model. Recent developments on this distribution (Huerta and Sansó 2007, Sang and Gelfand 2009) have focused on

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_49,

parameter estimation when temporal and spatial components are included in the model.

We construct confidence bands of GEV cumulative density functions (CDFs) in order to compare different pollution situations. Indeed, in GEV distributions, confidence intervals on parameters are difficult to be considered jointly (Hurairah et al. 2006): separate confidence intervals for each parameter fail to represent the real differences between CDF distributions. Extreme pollution situations are described using indices based on maxima, although other indices can be obtained using aggregating functions based on percentiles, in order to summarize less critical situations. In such cases, an analytically tractable distribution is not immediately obtainable, while comparisons in terms of CDFs are very difficult to achieve. However, CDFs and their confidence bands can be suitably estimated using non-parametric methods. For extreme pollution indices, both parametric and non-parametric methods can be employed to construct CDFs and their confidence bands. The loss in precision in the case of non-parametric methods is compensated for by the fact that they represent the natural starting point for extending the estimation of CDFs to indices based on percentiles. This paper is organized as follows: the next section sketches the probability distributions of air quality indices; Section 3 illustrates two different approaches to the probabilistic assessment of index distribution: the parametric GEV distribution approach and a non-parametric approach based on the Moving Block Bootstrap. The results of the two approaches are compared; while the fourth and final section offers some concluding remarks.

2 Probability Distributions of Air Quality Indices

In a previous study (Bruno and Cocchi 2002) we proposed a general class of air quality indices, capable of taking into consideration both different orders of aggregation (starting either from pollutants or from monitoring sites) and different aggregating functions. Indeed, as was also pointed out in Bodnar et al. (2008), the joint consideration of several indices reveals various different features of pollution. When the focus is on extreme cases, the index is constructed by selecting the maxima in the two dimensions at time t. The first step in the process of aggregation is: $X_{M(j)}(t) = Max_k(X_{q(kj)}(t))$, representing the synthesis of K monitoring sites, where $j = 1, ..., J$ runs among pollutants, while subscript M indicates that the chosen aggregating function is the maximum. Subscript q denotes the temporal synthesis (not necessarily the daily scale). The last step in the hierarchy consists in computing $I_{M(M)}(t)$, representing the second selection among the maxima:

$$I_{M(M)}(t) = Max_j\, f(X_{M(j)}(t)) \tag{1}$$

where f is the standardizing function required for aggregation across pollutants. The higher the index, the more critical the pollution. Expression (1) summarizes the hierarchical construction of the index. The hierarchical selection of maxima is the

special case of invariance with respect to the order of aggregation, and corresponds to measuring the most extreme pollution situations. Different pollution situations can also be compared by considering the uncertainty associated with the index.

The GEV probability distribution is suitable for a wide class of stationary processes, including sequences of time dependent data (Coles 2001, Chap. 5):

$$G(z) = exp[-(1 + \xi \frac{(z - \mu)}{\sigma})^{-1/\xi}], \tag{2}$$

characterized by μ, the location parameter, σ, the scale parameter, and ξ, the shape parameter. It is suitable for summarizing the extreme pollution situations described by index (1). The estimated GEV parameters and their standard errors can be obtained by means of the maximum likelihood (ML) method. The ML parameter estimates can be used to obtain the analytical expression of the CDF for each series in question. In Hurairah et al. (2006), the emphasis is on the difficulty of making inferences regarding CDF variability from separate parameter confidence intervals. Comparisons between these confidence intervals create certain difficulties, and the real differences between distributions do not emerge.

An alternative method of constructing CDFs that does not postulate special assumptions regarding index probability distribution, is the Moving Block Bootstrap (MBB, Künsch 1989, Liu and Singh 1992); this is a re-sampling method for assigning measures of accuracy to statistical estimates when observations are finite time-series of correlated data. Blocks of a fixed length are determined (here the block length is chosen by following Mignani and Rosa 1995), and blocks are randomly selected with replacement. The union of all the sampled blocks constitutes the final sample. The underlying idea is that if block length is suitably chosen, observations belonging to different blocks are nearly independent, while the correlation present in those observations forming each block is retained.

The confidence bands obtained for the CDF (in the parametric and non-parametric frameworks) are not analytically computable, but can be determined by means of simulation. The tool for comparing the simulated confidence bands is based on MonteCarlo tests (Lixing 2005). A set of 1,000 MonteCarlo replications of daily data from each estimated CDF, and their respective confidence bands, have been constructed on the basis of the simulated samples.

3 Probabilistic Assessment of Air Quality Index Distributions

We developed the assessment of differences in pollution by comparing confidence intervals for the whole CDFs. Daily values of index (1) were calculated for variously overlapping time periods in three cities (Bologna, Rome and Palermo), where pollutants O_3, Benzene, PM_{10}, CO, SO_2 and NO_2 were measured. The maximum likelihood estimates of GEV parameters with their standard errors are summarized

Table 1 Maximum likelihood estimation of GEV parameters (with their s.e.)

City	Years	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$
Bologna	2001–2002	80.80 (1.01)	24.75 (0.74)	0.08 (0.02)
Rome		103.21 (1.32)	32.09 (0.97)	0.02 (0.03)
Bologna	2002–2003	84.24 (1.11)	26.73 (0.86)	0.15 (0.03)
Palermo		101.32 (0.99)	24.54 (0.72)	0.04 (0.02)
Palermo	2002	100.14 (1.52)	26.12 (1.12)	0.08 (0.03)
Rome		94.02 (1.79)	30.34 (1.33)	0.07 (0.04)

Table 2 MonteCarlo CI on the distributions of differences between percentiles (in column) for the GEV distribution and the MBB setup (along separate rows)

		CI	P50	P75	P80	P85	P90	P95
Bologna-Rome	GEV	5%	−28.86	−35.48	−37.32	−39.65	−43.34	−48.80
2001–2002 (Case 1)		95%	−21.33	−25.10	−25.71	−26.15	−26.50	−24.14
	MBB	5%	−41.33	−50.32	−53.33	−57.66	−62.94	−69.27
		95%	−11.42	−14.65	−14.73	−15.15	−13.55	−0.66
Bologna-Palermo	GEV	5%	−19.52	−16.99	−15.69	−14.33	−12.16	−8.58
2002–2003 (Case 2)		95%	−12.94	−6.98	−4.38	−1.23	4.97	19.97
	MBB	5%	−26.03	−22.99	−23.97	−28.15	−27.38	−23.63
		95%	−4.69	5.34	12.31	20.27	26.68	32.66
Palermo-Rome	GEV	5%	−9.58	−8.52	−8.37	−8.49	−8.60	−11.80
2002 (Case 3)		95%	0.43	6.66	8.62	11.32	15.42	24.37
	MBB	5%	−14.35	−23.40	−26.17	−30.31	−34.25	−39.02
		95%	15.76	24.52	25.76	28.88	34.49	52.90

in Table 1. Since the aim is to compare pairs of pollution situations, the common overlapping periods have been isolated for each comparison.

As mentioned before, separate confidence intervals for parameters do not give a clear idea of the differences between distributions.

We suggest that confidence bands, based on the whole CDFs, be constructed by adopting two different approaches: the first approach assumes a parametric distribution for air quality indices, whereas the second approach is completely free of all assumptions. The overlapping of confidence bands is the criterion we propose to establish whether two distributions are significantly different or not.

We compare CDFs by means of simulation: MonteCarlo tests (Lixing 2005) are evaluated over the entire range of percentiles of the distributions under comparison. Differences in distributions are assessed using MonteCarlo testing as follows. For each replication, the differences between any pair of distribution functions are summarized by the differences between their percentiles: we report results for 50, 75, 80, 85, 90 and 95 percentiles. For each percentile, 1,000 replications of the differences are employed to compute the bilateral 90% MonteCarlo confidence interval (CI). The three comparisons represent the three possible situations, as shown in Table 2 and Figs. 1–3: totally different CDFs (Case 1), CDFs that are different up to a benchmark value (Case 2), and undistinguishable CDFs (Case 3). In Table 2, for

each identified case, the first two rows show the results based on the GEV distribution, whereas the second two rows show the results based on the MBB approach. If the MonteCarlo tests give negative values for all percentiles, then the differences between distributions are significant, and one of the distributions always assumes the highest values for all percentiles (Case 1); if the percentiles give rise to MonteCarlo tests that are negative up to a specific value (benchmark) this means that the distributions are different up to such benchmark, and then tend to be undistinguishable (Case 2). After the benchmark, in fact, the MonteCarlo test shows that the value 0 is included between the extremes. The interpretation of the benchmark value will be discussed later. The last case (Case 3) consists in test results that include 0 for all percentiles. When this occurs, the CDFs are considered undistinguishable.

Figures 1–3 compare the pairs of CDFs in the three cases in question. In each figure, the left panel contains the confidence bands of the GEV distribution, whereas the right panel displays the confidence bands of the MBB simulated distribution. These confidence bands are always wider than those based on the GEV distribution, due to the lack of assumptions of the MBB method. Case 1 denotes the situation where the CDFs are clearly different: all 90% confidence intervals of the distributions of differences exhibit values that are significantly different from zero, as shown in Table 2. This is also confirmed by Fig. 1, where one of the CDFs exhibits values that can always be considered lower than those of the other. This occurs when comparing Bologna (lower pollution levels) with Rome in 2001–2002. A clear-cut situation is also apparent from the parameter confidence intervals in Table 1, and is not affected by the different amplitude of confidence bands in the two contexts. In Fig. 1, when the CDFs tend towards 1, the two confidence bands seem to overlap, although this is merely due to a scale effect.

The second case (Fig. 2) concerns percentiles that are significantly different up to a benchmark value: above it, i.e., when the most critical situations begin to

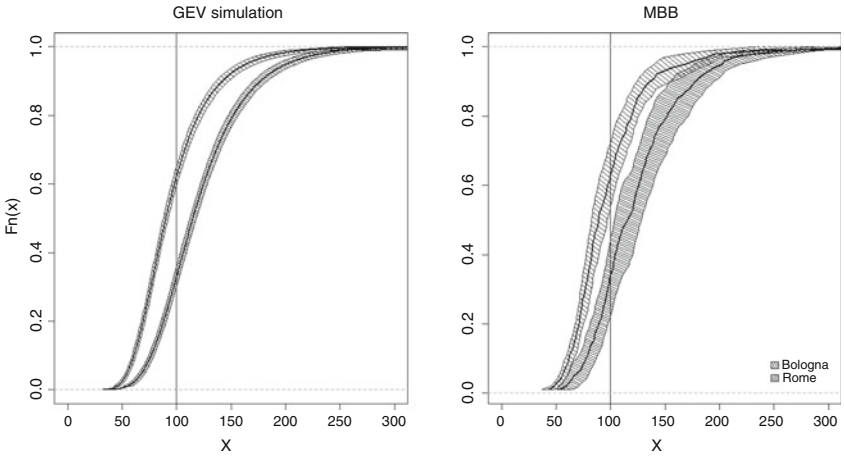


Fig. 1 Air quality indices CDFs with confidence bands: different CDFs (Case 1)

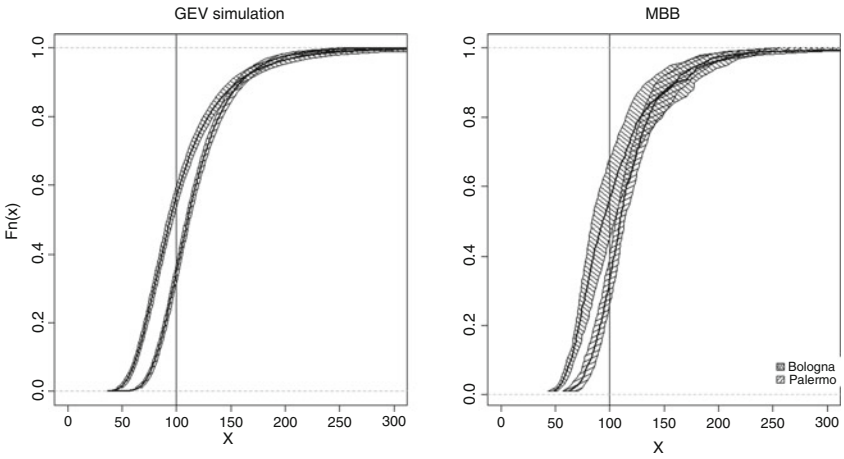


Fig. 2 Air quality indices CDFs with confidence bands: different up to a benchmark (Case 2)

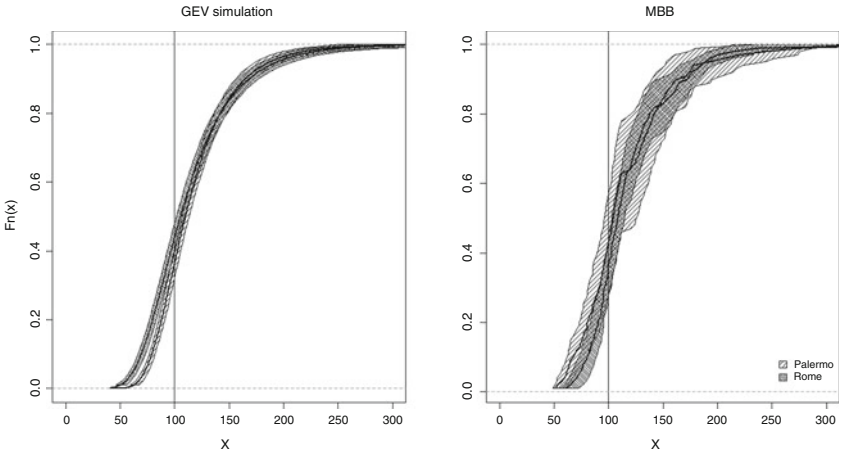


Fig. 3 Air quality indices CDFs with confidence bands: undistinguishable CDFs (Case 3)

emerge, the CDFs are not significantly different. This case is very interesting, since it enables us to identify the benchmark value below which a location shows a maximum pollution situation less serious than that of its competitor. For values below the benchmark, the critical pollution situations in the two locations are always separate: this is the case in the comparison between Palermo (higher values of pollution) and Bologna in 2002–2003, while the results shown in Table 1 are difficult to interpret. Furthermore, in Fig. 2 the differences in the amplitude of bands are due to differences in assumptions: they are more restrictive in the case of the first approach, corresponding to thinner confidence bands. The benchmark values are different: in the case of the GEV approach, the benchmark is the 87th percentile (corresponding

approximately to 147); whereas in the case of the MBB approach, the benchmark is the 67th percentile (corresponding approximately to 114).

The last case (Fig. 3) describes undistinguishable situations; it is difficult to reach this conclusion by looking only at the results of Table 1, which are difficult to interpret jointly.

3.1 Role and Interpretation of Benchmark Values

The benchmark represents the value beyond which the pairs of situations being compared are undistinguishable. Hence its interpretation refers to the thresholds established by law. When the authorities, such as the municipalities, try to reduce pollution, if the benchmark value of a comparison is higher than the critical value, analogous restrictions (to traffic or other sources of pollution) in both areas ought to be proposed. Consider Fig. 4 comparing air quality indices for the cities of Bologna and Palermo – which up until now have been considered jointly – with regard to two different years: 2002 in the left panel, and 2003 in the right panel. The benchmarks differ from one year to the other: for 2002 (*left panel*) the benchmark is 175, i.e., the 94th percentile, while for 2003 (*right panel*) the benchmark is 115, that is, the 60th percentile. In the first case, pollution should be considered different for most of the support of the CDFs: pollution reduction measures ought to be assessed and discussed essentially for the most polluted area. The case in 2003 is very different, since the benchmark, which corresponds to a percentile that is not particularly high, is very close to the critical 100 threshold separating good and poor air quality by law. The two areas ought to be considered as similar with regard to decisions aimed at reducing pollution.

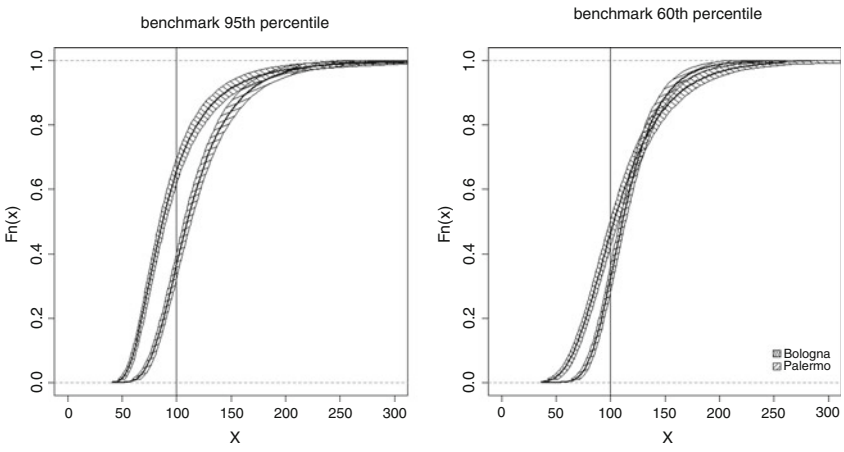


Fig. 4 Comparison of CDFs and their confidence bands with different benchmarks

4 Conclusions

We propose a way of comparing multi-pollutant air quality situations by considering synthetic air quality indices to be random quantities to which a CDF (parametric or not) can be associated, and which can be enriched by confidence bands. The tools proposed here are not only of scientific interest, but are also easy to compute, and can be adopted by policy makers and environmental organizations for ex-post assessment. The case study examined here would seem to weigh slightly in favour of the solution based on the GEV distribution, since the focus is on daily recorded maxima. Comparisons between distributions of air quality indices in the general class proposed in Bruno and Cocchi (2002), involve indices based on any percentile selections. The structures of such indices are very complex, and their probability distributions do not have easy to compute closed forms: the Moving Block Bootstrap approach appears preferable in these cases, where distributional assumptions are not available. Future analyses may want to compare CDFs for observed indices with a "theoretical" CDF obtained from the combined selection of thresholds for those pollutants included in the computation of the quality index. In this way, the method proposed here would change from being an ex-post evaluation, to a form of support to policy makers benefiting from statistical analysis.

Acknowledgements The research leading to this paper has been partially funded by a 2008 grant (Project n. 2008CEFF37-001, sector:13: Economics and Statistics) for research of national interest by the Italian Ministry of the University and Scientific and Technological Research.

References

Beirlant, J., Goegebeur, Y., Segers, J., & Teugels, J. (2004). *Statistics of extremes: Theory and applications*. Chichester: Wiley.

Bodnar, O., Cameletti, M., Fassó, A., & Schmid, W. (2008). Comparing air quality among Italy, Germany and Poland using BC indexes. *Atmospheric Environment*, *36*, 8412–8421.

Bruno, F., & Cocchi, D. (2002). A unified strategy for building simple air quality indices. *Environmetrics*, *13*, 243–261.

Bruno, F., & Cocchi, D. (2007). Recovering information from synthetic air quality indices. *Environmetrics*, *18*, 345–359.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. London: Springer-Verlag.

Huerta, G., & Sansó, B. (2007). Time-varying models for extreme values. *Environmental and Ecological Statistics*, *14*, 285–299.

Hurairah, A., Ibrahim, N. A., Daud, I. B., & Haron, K., (2006). Approximate confidence interval for the new extreme value distribution. *Engineering Computations*, *23*, 139–153.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, *17*, 1217–1241.

Lagona, F., & Maruotti, A. (2009). *A hidden Markov model for pollutants exceedances counts*. Graspa Working Paper nr. 33,.

Lee, D., Ferguson, C., & Scott, E. M. (2009). Air quality indicators in health studies. In S. Ingrassia & R. Rocci (Eds.), *Classification and data analysis* (pp. 217–220). Padova: Cleup.

Liu, R. Y., & Singh K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In R. Lepage & L. Billard (Eds.), *Exploring the limits of bootstrap*. New York: Wiley.

Lixing, Z. (2005). Nonparametric Monte Carlo tests and their applications. *Lecture notes in statistics* (Vol. 182). Berlin: Springer-Verlag.

Mignani, S., & Rosa, R. (1995). The moving block bootstrap to assess the accuracy of statistical estimates in Ising model simulations. *Computer Physics Communications*, *92*, 203–213.

Ruggeri, M., Plaia, A., & Bondí, A. L. (2009). Aggregate air pollution indices: A new proposal. In S. Ingrassia & R. Rocci (Eds.), *Classification and Data analysis* (pp. 221–224). Padova: Cleup.

Sang, H., & Gelfand, A. E. (2009). Hierarchical modeling for extreme values observed over space and time. *Environmental and Ecological Statistics*, *16*, 407–426.

Comparing Air Quality Indices Aggregated by Pollutant

Mariantonietta Ruggieri and Antonella Plaia

Abstract In this paper a new aggregate Air Quality Index (AQI) useful for describing the global air pollution situation for a given area is proposed. The index, unlike most of currently used AQIs, takes into account the combined effects of all the considered pollutants to human health. Its good performance, tested by means of a simulation plan, is confirmed by a comparison with two other indices proposed in the literature, one of which is based on the Relative Risk of daily mortality, considering an application to real data.

1 Introduction

Many epidemiological studies have recently shown a relevant association between adverse health outcomes and air pollution exposure. In the last years hospital admissions for respiratory and cardiovascular diseases have been increasing, causing morbidity and mortality as well as increased costs for National Health Services. For this reason, both people and policy-makers have been interested in AQIs; the knowledge of air pollution levels allows us to plan abatement strategies and precautionary measures to safeguard environment and citizens health.

Several countries of the world provide an AQI, but there is not a unique and internationally accepted methodology for constructing it. Although all the proposed AQIs aim to achieve the same objective, they may differ from one country to another in several aspects. The most used and up-to-date current AQI systems are based on US EPA (Environmental Protection Agency) AQI (http://www.epa.gov/ebtpages/airairquality.html). In Europe, the CAQI (Common Air Quality Index), implemented by CITEAIR, a project aiming at supporting European cities in pollution issues (Elshout et al. 2008), is computed (as EPA AQI) by linear interpolation between the class borders according to the grid in a table. Nevertheless, the breakpoints of the used tables, as well as the descriptors of each category, the air quality standards (AQS) and the functions dictated by the national agencies to aggregate hourly values

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_50,

daily, usually change from country to country. For example, the threshold value for $PM_{10}(24h)$ in the US system is about three times higher than the UK value. However, the main shortcoming of the described systems is that they do not take into account the adverse effects due to the coexistence of all pollutants; indeed, the final index is usually represented by the maximum value among the sub-indices computed for each pollutant, providing an underestimate of air pollution levels. Such a problem contrasts with some important properties that a good AQI should satisfy: it should not be ecliptic, that is it should not indicate a highly polluted air as less polluted, giving a false sense of security. Likewise, a good AQI should not be ambiguous, declaring an alarm situation, when it is unnecessary. In this paper we try to find an acceptable compromise between the two extreme situations of eclipsicity and ambiguity. To do this, we follow a data-driven approach, the most common in the literature, although a model based approach, modelling the data by a stochastic process, could be followed. This last approach has the advantage to account for missing data and to allow forecasting of future values for the AQI; however, it usually lacks simplicity and interpretability, two other basic requirements for a good AQI. In Sect. 2, since air pollution data are collected according to different dimensions, we describe how to aggregate them and, in particular, we emphasize the importance of the choice of the standardization process when the different pollutants have to be combined. Our index is presented in Sect. 3; a simulation study is performed in order to test its validity and to choose the most appropriate value for the parameter it depends on. In Sect. 4 our AQI, the index proposed by Bruno and Cocchi (2002) and the index proposed by Cairncross et al. (2007) are compared and applied to real data. Finally, in Sect. 5, some conclusions are drawn and some topics for further development are presented.

2 Aggregating and Standardizing Data

Air pollution data are usually collected according to time, space and type of pollutant. The three different dimensions have to be aggregated in order to provide a unique final index. The value of the synthetic index depends on the choice of the aggregating functions and the order of aggregation (Bruno and Cocchi 2002). Usually, data are first aggregated by time, according to the guidelines provided by national agencies. The functions here used to aggregate hourly values at each site for each pollutant are reported in Table 1. By considering a whole year dataset containing hourly data regarding K pollutants recorded at J monitoring sites, after the application of such functions, an $I \times J \times K$ array is obtained ($I = 365$ days).

In order to compare different pollutants which have different measurement units or order of magnitude, the simplest standardization may be performed by dividing the pollutant concentration by a threshold value, i.e. the maximum permissible concentration according to the directives (Bruno and Cocchi 2002). Such a procedure is

Table 1 Breakpoints of the AQI_k ($\mu g\, m^{-3}$ for all pollutants and $mg\, m^{-3}$for CO)

Pollution category	AQI_k	PM_{10} 24h	NO_2 1h	CO 8h	SO_2 24h	O_3 8h
Unhealthy	85–100	238–500	950–1900	15.5–30	500–1000	223–500
Unhealthy for sensitive groups	70–85	144–238	400–950	11.6–15.5	250–500	180–223
Moderate pollution	50–70	50–144	200–400	10–11.6	125–250	120–180
Low pollution	25–50	20–50	40–200	4–10	20–125	65–120
Good quality	0–25	0–20	0–40	0–4	0–20	0–65

adopted in Italy by some Regional Agencies for Environmental Protection (ARPA) and by some cities, interested in providing AQIs. Recently, it has been demonstrated that air pollution may cause long-term as much as short-term adverse health responses, so comparing the pollutant measured concentrations only with the guidelines established by the local agencies may be misleading, as it gives little emphasis to possible chronic health effects and to damages caused by air pollution on vegetation, animals and monuments, occurring over long time periods. Actually, there are no threshold values below which no adverse health response may be expected: it is known, for example, that some carcinogenic substances may even have latency period of years or decades. An alternative and more valid approach can be represented by an index measurable on a numerical scale, divided into different categories identified by descriptors, that quantify the impact of a mixture of air pollutants with respect to the well-being and health of people. For this purpose, the standardizing transformation by linear interpolation, used by US EPA and by CITEAIR, is to be preferred, as classes for low concentrations of a pollutant are also considered (long-time effects). It is based on the following segmented linear function, introduced by Ott and Hunt (1976):

$$AQI_k = \frac{I_{k,\,H} - I_{k,\,L}}{BP_{k,\,H} - BP_{k,\,L}}(C_k - BP_{k,\,L}) + I_{k,\,L} \tag{1}$$

where:

- AQI_k is the sub-index for pollutant k;
- C_k is the concentration (daily synthesis) of the pollutant k;
- $BP_{k,\,H}$ and $BP_{k,\,L}$ are the breakpoints $\geq C_k$ or $\leq C_k$, respectively;
- $I_{k,\,H}$ and $I_{k,\,L}$ are the AQI_k values corresponding to $BP_{k,\,H}$ and $BP_{k,\,L}$, respectively.

The breakpoints considered here are established according to EU standards and directives (European Community 2008) or come from epidemiological studies on single pollutants (Murena 2004), with a range [0,100] for each sub-index AQI_k (see Table 1). An AQI_k value of 50 corresponds to EU threshold values, so that the alarm level is a value greater than 50, meaning that a pollutant is in a dangerous range for human health on a given day. Obviously, bounds of Table 1 can be adapted according to directives dictated by the country for which air quality is to be assessed.

3 The Proposed AQI and the Simulation Study

After the aggregation over time and the standardizing transformation, the aggregating function among pollutants proposed by Swamee and Tyagi (1999), and then used by Kyrkilis et al. (2007), is considered:

$$I_\rho = (\sum_{k=1}^{K}(AQI_k)^\rho)^{\frac{1}{\rho}}. \tag{2}$$

The constant ρ varies in $[1, +\infty[$. Equal weights are given to the K sub-indices since a standardization is performed before computing I_ρ.

It can be noted that function (2), if multiplied by $(\frac{1}{K})^{\frac{1}{\rho}}$, is a subset of the family of power means that, as is known, is defined for ρ values ranging in $]-\infty, +\infty[$. The greater the parameter ρ, the greater the contribution of the largest values to the values of the power mean: if we increase ρ infinitely, the value of the power mean approaches the maximum value; in contrast if we decrease ρ infinitely, then the value of the power mean approaches the minimum value. Anyway, the family of power means is not considered here, as a mean would nullify any additive effect among pollutants.

To consider $\rho = 1$ in (2) means to sum all the computed AQI_k, by assuming linearly additive effects among pollutants; in this case, I_ρ provides a possible overestimate of air pollution levels. Actually, although the combined effect of all pollutants have to be taken into account, it is also true that a correlation among pollutants occurs. Setting $\rho \to \infty$ means we consider the maximum among the single AQI_k, not accounting for the combined effects of the pollutants; in this case, I_ρ provides an underestimate of air pollution levels. As mentioned earlier, this last method is the most used (US EPA, CITEAIR, Italian AQIs). Since setting $\rho = 1$ or $\rho \to \infty$ might cause unnecessary alarm or false sense of security, respectively, a good compromise has to be found within the range $]1, \infty[$. To this end we carry out a study on ρ via a simulation study, investigating the behaviour of I_ρ for a number of simulated scenarios and different values of ρ. Since, according to Table 1, each sub-index AQI_k may fall in $T = 5$ different classes, labelled from 1 to 5 from the worst to the best, we consider all the possible scenarios obtained by the combination with repetition $C(T, K) = C(5, 5) = 126$ of T elements choose K:

- scenario 1: 1 1 1 1 1;
- scenario 2: 1 1 1 1 2;

  .....
- scenario 126: 5 5 5 5 5.

The ordering of the 126 scenarios changes according to the value of ρ. By assuming that all the values in each class are equally distributed, for each of the 126 scenarios, 5 random deviates from a Uniform density, with lower and upper limits corresponding to the class in which each pollutant falls, are generated. In Fig. 1 the five simulated sub-indices AQI_k, for each of the 126 scenarios, and the related I_ρ

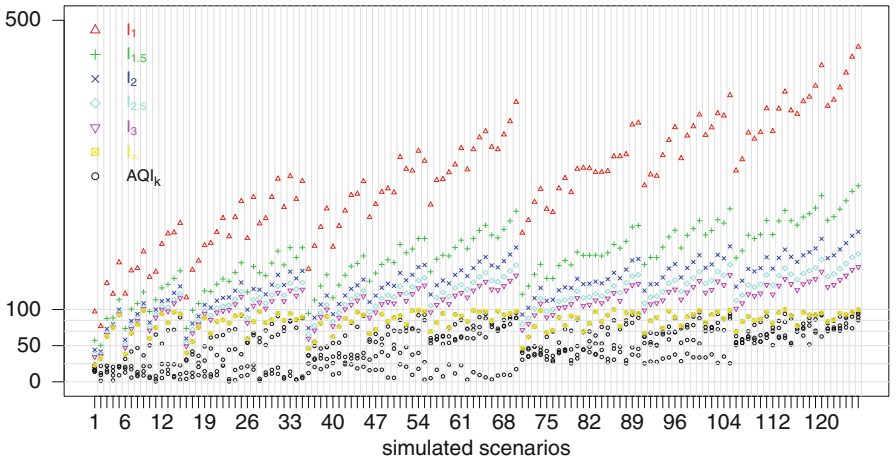


Fig. 1 Simulated AQI_k and I_ρ values computed on different possible scenarios

values, computed for different values of ρ ($\rho = 1, 1.5, 2, 2.5, 3, \infty$), are reported. Figure 1 suggests $\rho = 2$ as a suitable value: the index I_2 ($0 \leq I_2 \leq 223.61$) seems to include the effects of all the pollutants and, at the same time, minimizes both eclipsicity and ambiguity. Nevertheless, as pointed out also by Khanna (2000), the most appropriate value of ρ depends on the air quality: the more the air is polluted, the less strong the additive effects among pollutants are . Therefore, a value of ρ greater than 2 may be considered when at least one AQI_k falls in a high class, but this is not the case of our data set, and more generally of data recorded in Italy, falling at most in class 3 of Table 1 (moderate pollution).

4 Comparison Among Indices: An Application to Real Data

A dataset consisting of the values of the five main pollutants (NO_2, CO, PM_{10}, O_3, SO_2), recorded in three Italian cities (Padova, Palermo and Torino) on five different days of three different months (May, August and December) during 2008, is analyzed here. Since no spatial aggregation is considered in this paper, a single monitoring station, placed in the city centre, is selected for each city; it is not representative of the city itself, our purpose being to compare different indices and not different cities. In this example, no unreliable or missing data occur, although gaps are a crucial concern in air pollution data sets, due to measurement errors or misfunctions of instruments in the monitoring network. In a data-driven approach like this, they have to be replaced before computing any aggregating function; for this purpose, a simple average or more sophisticated imputation techniques may be considered, but this is not the aim of this paper.

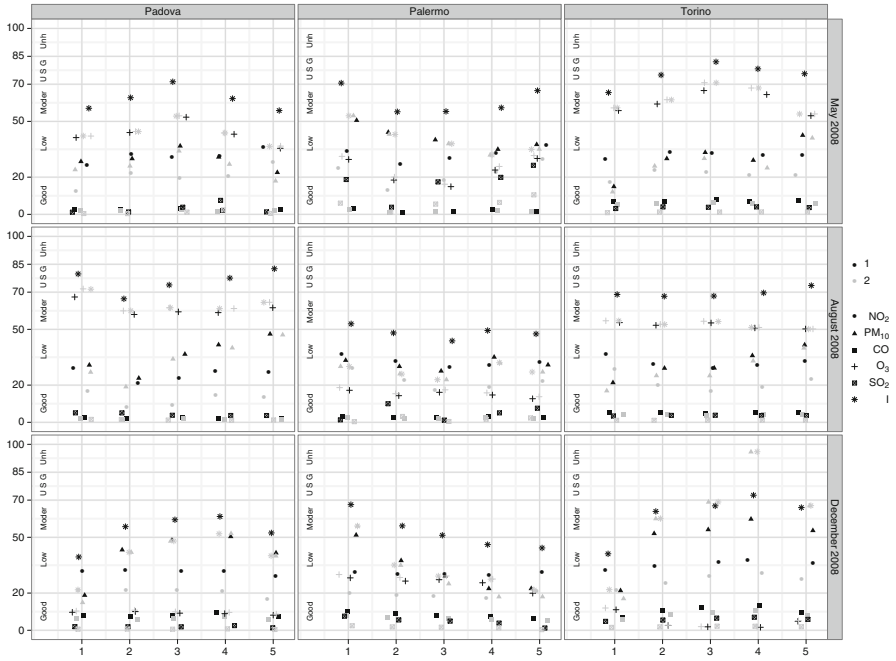


Fig. 2 Standardized concentrations, in black by linear interpolation (1) and in grey by threshold value (2), I_2 (*black star*) and I_∞ (*grey star*)

In Fig. 2 the indices I_2 (black star), computed on concentrations standardized by linear interpolation (labelled with 1), and I_∞ (grey star), computed on concentrations standardized by threshold value (labelled with 2), according to Bruno and Cocchi (2002), are compared. In standardization 2, the threshold values used for each pollutant are those corresponding to $AQI_k = 50$ in Table 1; moreover, the ratio between C_k and its corresponding threshold value is multiplied by 50 instead of 100, in order to make the two standardizations comparable, so that 50 is the value corresponding to the threshold value for both standardizations.

Figure 2 shows how the two different standardizations, and the related indices I_2 and I_∞, provide very different results.

With the help of Fig. 1, $\rho = 2$ appears as a reasonable value for the parameter, but, in order to verify and support this idea, together with the idea of using (2) as an aggregating function by pollutants, a comparison of our results with those of an alternative index, the API (Air Pollution Index), proposed by Cairncross et al. (2007), is reported. API evaluates the daily mortality risk related to the combined exposure to common air pollutants. A set of relative risk values of daily mortality RR_k is used to calculate sub-indices for each pollutant (see Cairncross et al. 2007, Table 4, pg 8449). The final API is the sum of the normalised values of the sub-indices:

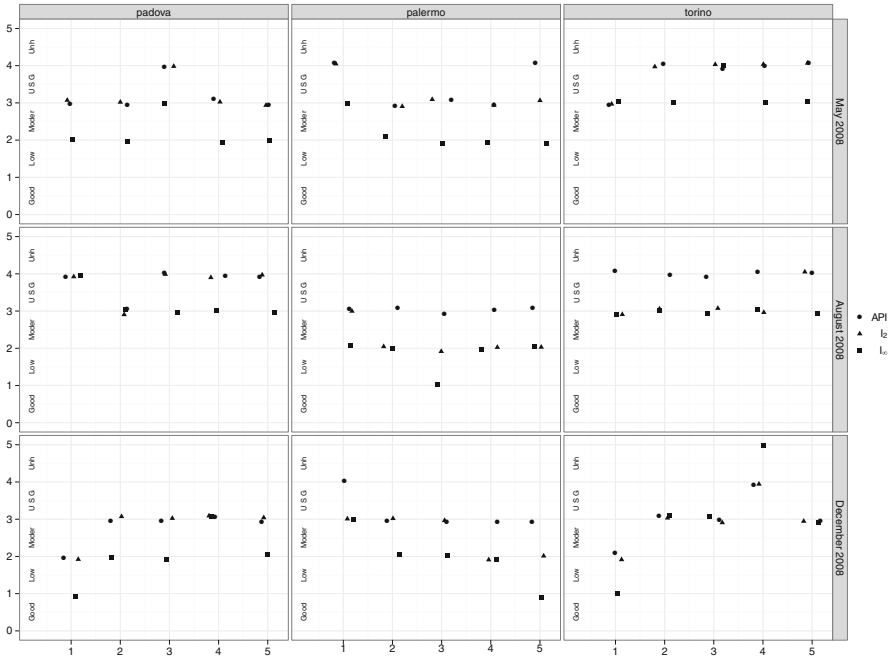


Fig. 3 Comparison among categorized I_2, I_∞ and API

$$API = \sum_k PSI_k, \tag{3}$$

where $PSI_k = a_k C_k$, with a_k a coefficient proportional to the incremental risk value $RR_k - 1$ (see Cairncross et al. 2007, Table 6, pg 8450). Each index value corresponds to a given daily mortality risk associated with the combined exposure to common air pollutants. In order to facilitate comparisons, we use the same RR_k values used by Cairncross et al. (2007), although updated values are available in the literature. Such values are computed under a procedure for health impact assessment in the EU and published by the World Health Organization.

Figure 3 shows the values obtained for I_2 and for API on the considered data set. To make the comparison easier, the class in which I_2 falls (according to Table 1) is reported, while for API, that assumes values in 0–10, the categorization reported by Cairncross et al. (2007) (see footnote, pg. 8450) is followed, that is: $0 \rightarrow 1$; $1-3 \rightarrow 2$; $4-6 \rightarrow 3$; $7-9 \rightarrow 4$; $10 \rightarrow 5$. In Fig. 3 the classes in which I_∞ falls are also reported. In the last case, standardization by threshold value is considered, according to Bruno and Cocchi (2002).

By observing Fig. 3, the difference between I_∞ and the other two indices is quite evident: I_2 and API fall more frequently in 'unhealthy' classes, while I_∞ almost always falls in lower classes, showing that I_∞ provides an underestimate of air pollution. The percentage of concordant pairs on the total, calculated between I_2

and API (73%), asserts the goodness of our index and validates the adequacy of the chosen value for ρ; it results to be better than the one calculated between I_∞ and API (18%).

5 Conclusions and Further Developments

In this paper a new aggregate AQI among pollutants is proposed. A simulation plan is used to choose the value of the parameter characterizing the family our index belongs to. A comparison with the index API, proposed by Cairncross et al. (2007), based on the Relative Risk of daily mortality, is performed in order to validate both the aggregating function and the parameter value chosen. The concordance (73%) between API Cairncross et al. (2007) and our index (I_2), greater than the one between API and I_∞ (Bruno and Cocchi 2002) (18%), validates the goodness of I_2. With reference to an aggregation among monitoring sites, which would produce a synthetic final index, describing air quality in a whole area (town/region) in a day, an approach based on Functional Principal Component Analysis is under study (Agrò et al. 2009).

References

Agrò, G., Di Salvo, F., Ruggieri, M., & Plaia, A. (2009). Air quality assessment via FPCA. In *TIES 2009 - the 20th Annual Conference of The International Environmetrics Society*. Bologna, July 5–9.

Bruno, F., & Cocchi, D. (2002). A unified strategy for building simple air quality indices. *Environmetrics*, *13*, 243–261.

Cairncross, E. K., John, J., Zunckel M. (2007). A novel air pollution index based on the relative risk of daily mortality associated with short-term exposure to common air pollutants. *Atmospheric Environment*, *41*, 8442–8454.

European Community. (2008). Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. *Official Journal*, *L 152*, 11/6/2008: 1–44.

Khanna, N. (2000). Measuring environmental quality: An index of pollution. *Ecological Economics*, *35*, 191–202.

Kyrkilis, G., Chaloulakou, A., & Kassomenos, P. A. (2007). Development of an aggregate AQI for an urban Mediterranean agglomeration: Relation to potential health effects. *Environment International*, *33*, 670–676.

Murena, F. (2004). Measuring air quality over large urban areas: development and application of an air pollution index at the urban area of Naples. *Atmospheric Environment*, *38*, 6195–6202.

Ott, W. R., & Hunt, W. F. (1976). A quantitative evaluation of the pollutant standards index. *Journal of the Air Pollution Control Association*, *26*, 1051–1054.

Swamee, P. K., & Tyagi, A. (1999). Formation of an air pollution index. *Journal Air and Waste Management Association*, *49*, 88–91.

van den Elshout, S., Leger, K., Nussio, F. (2008). Comparing urban air quality in Europe in real time: A review of existing air quality indices and the proposal of a common alternative. *Environment International*, *34*(5), 720–726.

Identifying Partitions of Genes and Tissue Samples in Microarray Data

Francesca Martella and Marco Alfò

Abstract An important challenge in microarray data analysis is the detection of genes which are differentially expressed across different types of experimental conditions. We provide a finite mixture model aimed at clustering genes and experimental conditions, where the partition of experimental conditions may be known or unknown. In particular, the idea is to adopt a finite mixture approach with mean/covariance reparameterization, where an explicit distinction among up-regulated genes, down-regulated genes, non-regulated genes (with respect to a reference probe) is made; moreover, within each of these groups genes that are differentially expressed between two or more types of experimental conditions may be identified.

1 Introduction

Microarray technology allows for the simultaneous measurement of thousands of gene expression levels within different types of experimental conditions. An important problem in this context is the detection of genes that are differentially expressed with respect to a known (or unknown) partition of experimental conditions; in a clinical context, the identification of such genes, which may be referred to as clinical markers, may improve diagnosis, early disease detection and lead to successful treatments. The most commonly used methods for identification of genes may be summarized into heuristic rules and model-based approaches. The simplest heuristic is based on the so called fold-change rule, where empirical cutoffs for gene expression changes between two classes are computed, without any assessment of statistical significance (data come with biological and technical variability). If we consider the use of robust statistical concepts, a primary way to detect genes which behave differently in a known two or multi-class problem is by looking at the standard t or F-test statistics, respectively; however, multiplicity problems may occur because thousands of hypotheses are tested simultaneously. Thus, adjusting for multiple testing has been proposed when assessing statistical significance of results (see e.g., Benjamini 1995, Efron and Tibshirani 2002, McLachlan et al. 2002). However,

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_51,

few papers have dealt with the problem of identification and validation of subsets of genes and experimental conditions (i.e., tissue samples), that is the identification of genes which behave differentially when given subsets of experimental conditions are concerned. For example, Martella et al. (2008) assume that gene profiles are drawn from a finite mixture distribution, with tissue (condition)-specific effects estimated through a parametric location/scale change driven by a known partition of experimental conditions. In this paper, we introduce an extension of this model to allow for the detection of differentially expressed genes in the case of unknown or known partitions of tissues. The proposed model is discussed using a large scale simulation study. The plan of the paper is as follows. In Sect. 2, we describe the model proposed by Martella et al. (2008). Section 3 discusses the hierarchical proposal. In Sect. 5, the proposed model is applied to simulated data sets. Last section entails concluding remarks and future research agenda.

2 Model-Based Biclustering

Martella et al. (2008) proposed a *biclustering* model for simultaneous clustering of rows and columns of a given data matrix, where the key idea is to approximate the data density by a mixture of Gaussian distributions with a particular component-specific covariance structure. More precisely, whereas a traditional mixture approach is used to define the gene clustering, they propose to use a binary and row stochastic matrix to represent column (i.e., tissues) partition. In details, let $\mathbf{y}_i$ be a J-dimensional vector representing the gene profile for the i-th gene over J experimental conditions ($i = 1, \ldots, n$). Let us assume that the observed J-dimensional vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ come from a mixture of a finite number, say K, of components in some unknown proportions $\pi_1, \ldots, \pi_k$, which sum to one. Conditional on the k-th component of the mixture, $\mathbf{y}_i$ is specified by a factor analysis model as follows:

$$\mathbf{y}_{ik} = \boldsymbol{\mu}_k + \mathbf{V}_k \mathbf{u}_{ik} + \mathbf{e}_{ik} \tag{1}$$

where $\boldsymbol{\mu}_k$ is the J-dimensional component-specific mean vector, $\mathbf{V}_k = \{v_{jl}\}$ ($j = 1, \ldots, J$, $l = 1, \ldots, Q_k$) represents the column-cluster membership, $\mathbf{u}_{ik}$ is a Q_k-dimensional ($Q_k < J$) vector of component-specific latent variables (factors), which are assumed to be i.i.d. draws from $N(\mathbf{0}, \mathbf{I}_{Q_k})$, and $\mathbf{I}_{Q_k}$ denotes the $Q_k \times Q_k$ identity matrix. Furthermore, $\mathbf{e}_{ik}$ are i.i.d. Gaussian component-specific random variables with mean $\mathbf{0}$ and covariance matrix $\mathbf{D}_k = diag(\sigma^2_{1k}, \ldots, \sigma^2_{Jk})$, that are assumed to be independent of $\mathbf{u}_{ik}$. It is worth noticing that this model leads to a block diagonal component-specific covariance matrix for $\mathbf{y}_i$ ($i = 1, \ldots, n$), which keeps with the microarray problems where we assume that observations are homogeneous (with small within group variance) only under limited block of conditions, which, therefore, are highly correlated. Model parameters are estimated through a maximum likelihood approach using an Alternating Expectation Conditional Maximization (AECM) algorithm (e.g., Meng and Van Dyk 1997). Finally, we should

mention that this model may be seen as a special case of a more general model; the mixture of structural equation models (mixture SEM) developed in the psychometrics context (Arminger and Stein 1997). By having this general framework as a starting point, we could extend the model proposed by Martella et al. (2008) relaxing some assumptions. For example, we could assume different specifications for the component-specific covariance matrix of $\mathbf{u}_{ik}$.

3 Extension to a More Flexible Cluster-Specific Distributions

A potential limitation of the model proposed in Martella et al. (2008) is represented by the potential non-Gaussianity of component-specific densities, which is frequently encountered in practice. In the current biological problem, the assumption of Gaussianity for the cluster density may lead to wrong conclusions whenever genes are on the border line. In most empirical cases, however, it is difficult to decide which parametric distribution is suitable to characterize a cluster. Therefore, it may be wise to model the intra-cluster gene specific extra Gaussian departures through a more flexible family of distributions. Here, we adopt the hierarchical mixture model described in Li (2005) and Vermunt (2003). To discuss this hierarchical extension, we need to extend the standard notation for mixture model considering an (unobserved) extra-level. Let us start assuming that $\mathbf{y}_i$ belongs to one of K clusters and that each cluster is composed by T_k components ($i = 1, \ldots, n$; $k = 1, \ldots, K$); the latter are used to model potential extra-Gaussian variation in cluster-specific densities. The adopted multilevel structure implies that the marginal density of $\mathbf{y}_i$ ($i = 1, \ldots, n$) can be written as follows:

$$f(\mathbf{y}_i|\boldsymbol{\phi}) = \sum_{k=1}^{K} \pi_k \sum_{t=1}^{T_k} \pi_{t|k} N(\mathbf{y}_i; \boldsymbol{\mu}_{t|k}, \boldsymbol{\Sigma}_{t|k}) \tag{2}$$

where $\boldsymbol{\phi}$ represents the vector of model parameters, π_k and $\pi_{t|k}$ the prior probabilities corresponding to second and first level components (with $0 \leq \pi_k \leq 1$, $\sum_{k=1}^{K} \pi_k = 1$ and $0 \leq \pi_{t|k} \leq 1$, $\sum_{t=1}^{T_k} \pi_{t|k} = 1$), $\boldsymbol{\mu}_{t|k}$ and $\boldsymbol{\Sigma}_{t|k}$ are the J-dimensional component-specific mean vectors and the corresponding $(J \times J)$ covariance matrices. In particular, conditional on the first level t-th component membership, we assume that $\boldsymbol{\mu}_{t|k}$ may be parameterized as follows

$$\boldsymbol{\mu}_{t|k} = \boldsymbol{\mu}_k + \mathbf{V}_{t|k}\mathbf{u}_{it|k} \tag{3}$$

where $\boldsymbol{\mu}_k$ represents a column-wise constant J-dimensional cluster-specific vector, with $\boldsymbol{\mu}_k = \alpha_k \mathbf{1}_J$, with $\alpha_k \in \mathrm{IR}$, $\mathbf{V}_{t|k}$ is the $J \times Q_{t|k}$ component-specific binary row stochastic matrix, while the $Q_{t|k}$-dimensional latent factors $\mathbf{u}_{t|k}$ project the component-specific mean deviations $(\boldsymbol{\mu}_{t|k} - \boldsymbol{\mu}_k)$ onto a low dimensional space $(Q_{t|k} < J)$, $t = 1, \ldots, T_k$ and $k = 1, \ldots, K$. In particular, the term $\mathbf{V}_{t|k}\mathbf{u}_{it|k}$

represents the deviations from the cluster-specific mean $\boldsymbol{\mu}_k$ due to the $t|k$-th tissue component membership. ML estimates can be achieved by means of an EM algorithm, where, due to the high dimensionality of the estimation problem, we have to use an upward-downward type algorithm (see e.g., Pearl 1988). As in standard EM algorithms, the observed data likelihood function does not decrease at each iteration and the sequence converges to a local optimum. There are a variety of heuristic approaches to escape from local maxima such as using several different random starts. Finally, the i-th gene ($i = 1, \ldots, n$) is allocated to the t-th component within the k-th cluster corresponding to the largest posterior joint probability, $w_{ik}w_{it|k}$ ($t = 1, \ldots, T_k, k = 1, \ldots, K$), while the j-th tissue sample is allocated to the l-th cluster ($l = 1, \ldots, Q_{t|k}$) by using the elements in the matrix $\mathbf{V}_{t|k}$.

It is well known from previous works (e.g., Willse and Boik 1999; Hastie and Tibshirani 1996; Di Zio et al. 2005) that a drawback of hierarchical mixture models is the potential non-identifiability. In fact, without additional constraints, the hierarchical mixture model reduces to a standard mixture model with a number of components equal to the sum of first level mixture components (Li 2005). This problem is strictly connected with the problem of merging components in a standard mixture model discussed in several recent proposals (see Tantrum et al. 2003; Li 2005; Baudry et al. 2008; Henning 2009). In fact, in order to ensure identifiability of the proposed model, we would need to fix the number of clusters we are interested in; or, in other words, to decide whether some components should be merged in order to interpret their union as a cluster. However, since the concept of cluster has not a unique definition, an optimal assumption does not hold for all practical situations. For example, we may look for a cluster with high variance of a particular variable without necessarily differing too much from another more homogeneous cluster on average; or viceversa, we may look for clusters with low variances but very far from each others. Therefore, in order to solve our problem, we need to focus on a specific practical situation. Here, we focus on cDNA microarray measured on different tissue samples, where the generic element under study is the relative gene expression value measured with respect to a given reference probe. In practice investigators are often interested in a three-cluster partition of genes; i.e., in distinguishing among up-regulated genes, down-regulated genes, non-regulated genes (with respect to the reference). Thus, we assume that the first cluster includes up-regulated genes ($\alpha_1 > 0$), the second cluster includes non-regulated genes ($\alpha_2 = 0$) and the third cluster includes the down-regulated genes ($\alpha_3 < 0$). In particular, we would like to point out that identifiability is guaranteed by fixing a priori the number of gene clusters ($K = 3$) and using ordering constraints on the cluster-specific means ($\boldsymbol{\mu}_1 > \boldsymbol{\mu}_2 > \boldsymbol{\mu}_3$). Finally, within each of these clusters, genes that are differentially expressed between two or more types of tissue samples (known or unknown) may be identified thanks to the hierarchical structure of the model. We would like to emphasize this model may help not only in distinguishing between clusters (second level components) and components (first level components), but also in explicitly capturing biologically meaningful differences of the obtained partitions.

4 Simulation Study

In this section, we present a simulation study to give an idea about the performance of the proposed model. We simulated 100 data sets, each contains 3,000 genes profiles for 10 abnormal and 20 normal tissue samples. The data set is built on $K = 3$ clusters: the biggest cluster includes non-regulated genes ($\pi_2 = 0.79$), while the other two clusters represent up ($\pi_1 = 0.12$) and down-regulated ($\pi_3 = 0.09$) genes. Among the up-regulated genes, we assume to have 70 differentially expressed genes $\pi_{1|1} = 0.19$ while the remaining are assumed to be not differentially expressed ($\pi_{2|1} = 0.81$). Other parameters of the setting are the following: $\boldsymbol{\mu}_1 = 3\mathbf{1}_J$, $\boldsymbol{\mu}_2 = 0\mathbf{1}_J$, $\boldsymbol{\mu}_3 = -5\mathbf{1}_J$, $\mathbf{u}_{1|1} = \sum_{i=1}^{n_{1|1}} \mathbf{u}_{i1|1}/n_{1|1} = [-5, 2]$, $\mathbf{u}_{2|1} = \sum_{i=1}^{n_{2|1}} \mathbf{u}_{i2|1}/n_{2|1} = 3$, with $n_{t|k}$ number of genes within the t-th component in the k-th cluster ($t|k = 1|1, 2|1$), sum($\mathbf{V}_{1|1}$) = [20, 10], sum($\mathbf{V}_{2|1}$) = [30], and a constant and diagonal error matrix given by $\mathbf{D}_{1|1} = \mathbf{D}_{2|1} = \ldots \mathbf{D}_{1|3} = \ldots = diag(0.52, \ldots, 0.43)$.

The performance of the proposed model has been evaluated in terms of recovering the true gene and tissue sample partitions; in particular, we measured the degree of agreement between the true (first and second level) gene and tissue partitions membership and the partitions estimated by the upward-downward algorithm by using three agreement indices: Modified Rand Index, Jaccard Index and Hubert Index (Milligan and Cooper 1986). In case of perfect agreement, the values of these three indices are equal to one. For all the simulated data sets, the algorithm starts from the solution of a standard K-means algorithm performed on both rows and columns of the data matrix. This procedure considerably improves the performance of the upward-downward algorithm requiring a significantly reduced computational time.

Moreover, we assess the performance of the model as a method for selecting differentially expressed genes from different prospectives, including FDR (the percentage of not differentially expressed genes among selected genes), FNDR (the percentage of differentially expressed genes among unselected genes), FPR (the percentage of selected genes among not differentially expressed genes), and FNR (the percentage of un-selected genes among differentially expressed genes). Usually it is challenging to estimate these error rates for real data sets, because we do not know whether a gene is differentially expressed. However, because we know the true gene membership for each simulated data set, we estimate error rates by directly comparing the true gene membership with the gene membership estimated by the proposed method. For example, FNR is estimated by the ratio of the number of unselected genes among differentially expressed genes to the total number of differentially expressed genes for a simulated data set. Of course, we estimated the error rates and agreement indices by averaging them over the 100 simulated datasets. The values of the estimated error rates and agreement indices are summarized in Fig. 1. As you can easily see, the model performs well in recovering the true partitions (all indices are above 0.8) and in term of error rates the model has small (below 0.2) estimated FDR, FNDR, FPR and FNR.

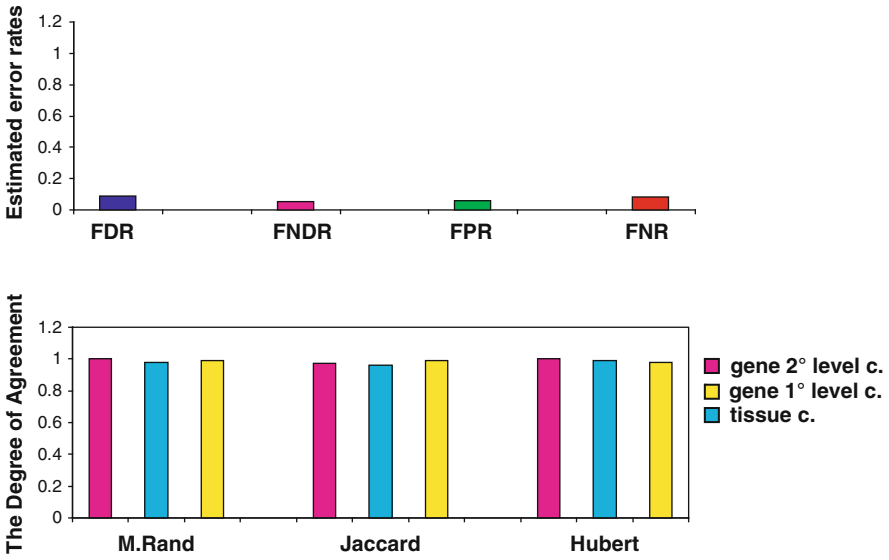


Fig. 1 Estimated error rates (*above*) and measures of degree of agreement (*below*)

Finally, in order to have an idea about the robustness of the proposed model with respect to the first level of the mixture model, that is if the model is able to discriminate between differentially and not differentially expressed genes, for each gene we compute the relative entropy given by:

$$Ent(\mathbf{y}_i) = -\frac{\sum_{k=1}^{K}\sum_{t=1}^{T_k} w_{it|k} w_{ik} log(w_{it|k} w_{ik})}{\frac{1}{K}\sum_{k=1}^{K} log(\frac{1}{K}\frac{1}{T_k})}. \quad (4)$$

Entropy increases when the ability of the model to discriminate between differentially and not differentially expressed genes decreases. Simple graphical summary of the relative entropy distributions for those genes selected as being differentially and not differentially expressed are represented in the Fig. 2. As can simply be observed, both distributions are centered around 0 although the distribution of the estimated not differentially expressed genes has slightly more variability maybe due to the high dimensionality of this cluster.

5 Conclusions and Future Research

In this paper, we propose a finite mixture model to identify partitions of genes and tissue samples in microarray data. We discuss a hierarchical extension of the finite mixture model proposed by Martella et al. (2008) to combine advantages of allowing gene clusters represented by mixture components and to identify which gene and tissue sample clusters are potentially meaningful. Our simulation results show that our proposal is encouraging and is worth to be applied on real data sets. Other

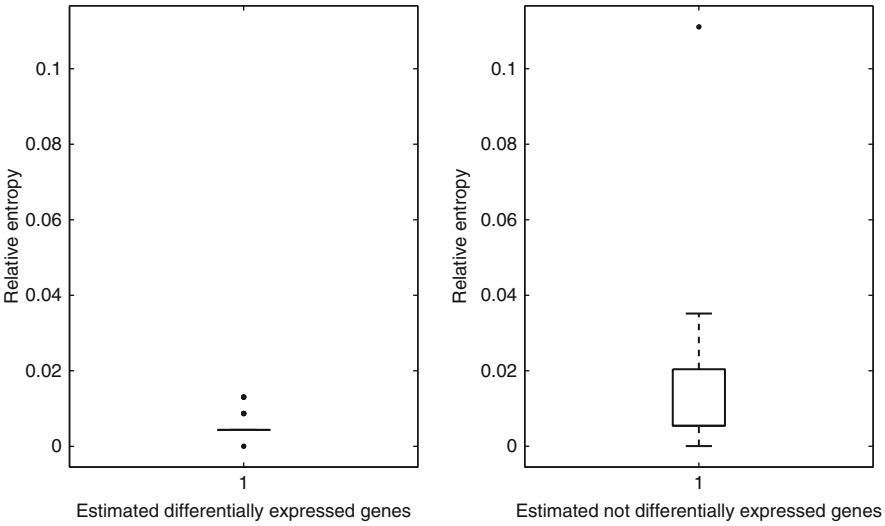


Fig. 2 Boxplot of the relative entropy distributions for the estimated differentially (*left*) and the estimated not differentially expressed genes (*right*)

simulation results (not presented here) show that in experimental designs where K can be considered known and T_k and $Q_{t|k}$ have to be estimated, BIC criteria performs well in selecting the true number of T_k and $Q_{t|k}$. Moreover, as far as the computational time is concerned, we noticed the following behavior:

- For fixed K and $Q_{t|k}$, the computational complexity increases with T_k, while the performance is almost unchanged.
- For fixed K and T_k, the computational complexity increases with $Q_{t|k}$, while the performance is almost unchanged.
- In more complex situations obtained by increasing K, T_k and $Q_{t|k}$ the upward-downward algorithm performs well in recovering the true partitions.

Various interesting future research lines are possible. The most straightforward ones would be: a simulation study with different error levels in the data to consider varying homogeneity levels within clusters, an application of the model to real cDNA data sets and a comparison of the performance of the proposed model with other benchmark models. More complex extensions to be mentioned are the possibility of using a soft tissue samples clustering and a variant of the model to allow for different gene expression design matrices.

References

Arminger, G., & Stein, P. (1997). Finite mixture of covariance structure models with regressors: loglikehood function, distance estimation, fit indices, and a complex example. *Sociological Methods and Research*, *26*, 148–182.

Baudry, J. K., Raftery, A. E., Celeux, G., Lo, K., & Gottardo, R. (2008). Combining mixture components for clustering. *Journal of Computational and Graphical Statistics, 19*(2), 332–353.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, *57*, 289–300.

Di Zio, M., Guarnera, U., & Luzi, O. (2005). Editing systematic unity measure errors through mixture modelling. *Survey Methodology*, *31*, 53–63.

Efron, B., & Tibshirani, R. (2002). Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, *23*, 70–86.

Hastie, T., & Tibshirani, R. (1996). Discriminant analysis by Gaussian mixtures, *Journal of the Royal Statistical Society. Series B*, *58*, 155–176.

Henning, C. (2009). Methods for merging Gaussian mixture components. *Advances in Data Analysis and Classification, 4*(1), 3–34.

Li, J. (2005). Clustering based on a multi-layer mixture model. *Journal of Computational and Graphical Statistics*, *14*(3), 547–568.

Martella, F., Alfò, M., & Vichi, M. (2008). Biclustering of gene expression data by an extension of mixtures of factor analyzers. *The International Journal of Biostatistics*, *4*, 1–3.

McLachlan, G. J., Bean, R. W., & Peel, D. (2002). A mixture model based approach to the clustering of microarray expression data. *Bioinformatics*, *18*(3), 413–422.

Meng, X. L., & Van Dyk, D. A. (1997). The EM algorithm -an old folk song sung to a fast new tune. *Journal of the Royal Statistical Society. Series B*, *59*, 511–567.

Milligan, G. W., & Cooper, M. C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, *21*, 441–458.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.

Tantrum, J., Murua, A., & Stuetzle, W. (2003). Assessment and pruning of hierarchical model based clustering. (KDD '03) *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM New York, NY, US, 197–205.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, *33*, 213–239.

Willse, A., & Boik, R. J. (1999). Identifiable finite mixtures of location models for clustering mixed-mode data. *Statistics and Computing*, *9*, 111–121.

Part XI
Analysis of Categorical Data

Assessing Balance of Categorical Covariates and Measuring Local Effects in Observational Studies

Furio Camillo and Ida D'Attoma

Abstract This paper presents a data driven approach that enables one to obtain a global measure of imbalance and to test it in a multivariate way. The main idea is based on the general framework of Partial Dependence Analysis (Daudin, 1981 J. J.) and thus of Conditional Multiple Correspondences Analysis (Escofier, B. (1988). Analyse des correpondances multiples conditionelle. La Revue de Modulad) as tools for investigating the dependence relationship between a set of observed categorical covariates (**X**) and an assignment-to-treatment indicator variable (**T**), in order to obtain a global imbalance measure (GI) according to their dependence structure. We propose the use of such measure within a strategy whose aim is to compute treatment effects by subgroups. A toy example is presented for illustrate the performance of this promising approach.

1 Introduction

Propensity Score (PS) and Propensity Score Matching have become popular methods of causal inference from observational data whose main characteristic is the lack of random assignment of units to treatment levels. The aim is to balance non-equivalent groups on observed pre-treatment covariates in order to reduce selection bias in the causal effect estimation. The success of PS and PS Matching in reducing bias mainly depends on balance criteria adopted. Balance concerns similarity in covariate distributions across treatment groups (Rubin 2001). As reported in Ho et al. (2007), it holds when **T** and **X** are unrelated such that $\tilde{p}(\mathbf{X} \mid \mathbf{T} = 1) = \tilde{p}(\mathbf{X} \mid \mathbf{T} = 0)$; where $\tilde{p}$ denotes the observed empirical density of data. Balance is commonly evaluated by conducting hypothesis testing. The standard practice involves the use of t-test for the difference in means for each continuous covariate or the χ^2 test for each categorical covariate. However, this practice starts to be widely criticized (Imai et al. 2006, Iacus et al. 2008). Despite others provide various approaches to deal with selection bias and test imbalance (Imai et al. 2006, Iacus et al. 2008), our attempt here is to provide a strategy that enables one to non-parametrically measure global imbalance and test it by preserving the multivariate nature of data. In

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_52,

particular, this paper points to a new multivariate approach (Camillo and D'Attoma 2010, D'Attoma and Camillo 2011) that uses Conditional Multiple Correspondence Analysis (MCA_{Cond}) (Escofier 1988) as a tool for investigating the dependence relationship between the **X** categorical covariates and the assignment-to-treatment indicator variable **T**. This paper is organized as follows. Section 2 introduces the global imbalance measure and the multivariate imbalance test. Section 3 illustrates how to measure unbiased treatment effects by subgroups. Section 4 shows using a toy example how the strategy works in practice and its performance in measuring treatment effects heterogeneity. Section 5 concludes.

2 Measuring and Testing Global Imbalance

In this section we propose a global imbalance measure based on the concept of between-groups inertia of a factorial predictor space.[1] When the dependence between **X** and **T** is out of control of researchers displaying the relationship among them on a factorial space represents a first step for discovering the hidden relationship. In fact, if dependence between **X** and **T** exists, any descriptive factorial analysis may exhibit this link. Thus, a conditional analysis could be useful in order to isolate the part of the variability[2] of the X-space due to the assignment mechanism. Conditioning applied to problems arising from the dependence between categorical covariates and an external categorical variable was first studied by Escofier (1988) with the resulting MCA_{Cond}. It consists in a factorial decomposition of the within- groups inertia after the between-groups inertia has been eliminated. Being the conditional decomposition model symmetric (Escofier 1988), two separated but equivalent spaces are obtained: a conditional unit space (R^P_{Cond}) and a conditional variable space (R^N_{Cond}). The conditional unit space is obtained by centering the t subspaces on its own center. In the dual space of categories, the conditional variable space is obtained by a two-stage procedure. In the first stage, it involves the projection of the column profile of the **X** (recoded in a disjunctive form) matrix onto R^T, the subspace generated by the t indicator variables; in the second stage, the structure induced by **T** is eliminated by making a projection on the space orthogonal to R^T. Conditioning is also related to a factorial analysis with reference to a model (Escofier 1984). In particular, it has been demonstrated (Escofier 1984) that the analysis of the divergency between a given frequency table and an independence model could be generalized to the analysis of the differences between a generic data matrix and a generic model. The latter is represented by a table which denotes the structure induced by **T** (Escofier 1988). Based on the Huygens inertia decomposition which decomposes the total inertia (I_T) of a considered space in within-groups (I_W) and

[1] The predictor space denotes the space generated by the pre-treatment covariates involved in the selection process.

[2] Here, we adopt the concept of inertia as a measure of association among categorical covariates.

between- groups (I_B), MCA_{Cond} could be considered as an *Intra Analysis* given that it does not consider the inertia (I_B) induced by the conditioning variable **T**. In order to obtain a global imbalance measure we consider as starting information the **X** matrix and the assignment-to-treatment indicator vector **T**. In particular, we assume to have enough information in the observed pre- treatment covariates: in this sense, we expect that the **X** matrix generally includes all pre-treatment variables associated with both the treatment assignment **T** and the outcome **Y**. Clearly, MCA_{Cond} enables one to quantify the dependence between **X** and **T** represented by the between-groups inertia (Estadella et al. 2005) defined as:

$$I_B = \frac{1}{Q} \sum_{t=1}^{T} \sum_{j=1}^{J_Q} \frac{b_{tj}^2}{k_{.t} k_{.j}} - 1 \tag{1}$$

where Q denotes the number of pre-treatment covariates considered, J_Q denotes the set of all categories of the Q variables considered, b_{tj} denotes the number of units with category $j \in J_Q$ in the treatment group $t \in T$, $k_{.t}$ denotes the group size $t \in T$ and $k_{.j}$ denotes the number of units with category $j \in J_Q$. We define the between-groups inertia (I_B) as the Global Imbalance Measure (GI) in data. The proposed measure varies in $[0, I_T]$. Perfect balance occurs when $I_B = 0$; whereas, perfect imbalance occurs when $I_W = 0$ and $I_B = I_T$ which indicates that the observed total variability of the **X**-space is completely due to the influence of conditioning (**T**). Once I_B has been obtained, we derive the Multivariate Imbalance Coefficient (MIC) (D'Attoma and Camillo, 2011), which is defined as one minus the ratio between the within-groups inertia relative to the total inertia:

$$MIC = 1 - \frac{I_W}{I_T} \tag{2}$$

where I_W denotes the inertia considered in the conditional analysis (MCA_{Cond})[3] and I_T denotes the inertia considered in the unconditional analysis (i.e., Multiple Correspondence Analysis). As a result, if the right covariates involved in the selection process have been considered, then I_B will represent the correct global imbalance measure and MIC effectively express its importance relative to the total inertia. Finally, we perform an hypothesis test (D'Attoma and Camillo, 2011) to determine the significance of the detected imbalance. We specify the null hypothesis of no dependence between **X** and **T** as:

$$H_0 : I_W = I_T \tag{3}$$

To establish an interval of plausible values for I_B under the null hypothesis , we use results obtained by Estadella et al. (Estadella et al. 2005), who have studied

[3] It has been demonstrated (Estadella et al. 2005) that after MCA_{Cond} the total inertia of the re-centered space (I_{Cond}) equals the original within-groups inertia (I_W)

the asymptotic distribution function of I_B. The distribution of the between-groups inertia was derived as $I_B \sim \frac{\chi^2_{(T-1)(J-1)}}{nQ}$, so that an α confidence interval can be obtained as:

$$I_B \in (0, \frac{\chi^2_{(T-1)(J-1),\alpha}}{nQ}) \tag{4}$$

Based on such result, if the I_B calculated on the specific data set under analysis is out the interval, then the null hypothesis is rejected and thus data unbalanced.

3 Measuring Heterogeneous Treatment Effects in Local Spaces

Combining the GI measure and the Imbalance Test, we propose a simple two-step approach for computing unbiased treatment effects in non-experimental data. First, we measure and test balance on the whole sample; then, if imbalance exists, we propose to measure local average treatment effects (Peck et al. 2010; Camillo and D'Attoma 2010). More precisely, a transition from the global predictor space to local predictor spaces is done in order to measure unbiased treatment effects . The name *global* indicates the entire predictor space and the term *local* refers to many local predictor spaces determined by specific combinations of covariates. Local spaces are found by means of a classification on factorial coordinates[4] which involves two steps: first, a low-dimensional representation of the X-space is obtained via Multiple Correspondence Analysis (MCA) (Benzécri 1973, Lebart et al. 1997); second, a Cluster Analysis (CA) is performed in order to identify homogeneous groups on the basis of the low-dimensional MCA coordinates.[5]

The use of CA is not new on evaluation field. Papers of Yoshikawa et al. (2001) and Gibson (2003) are examples of applications of cluster analysis in the field of experiments. Another application is in Peck (2005). The author proposes using cluster analysis to identify subgroups within experimental data, with the aim of understanding variation in program impacts that accrues heterogeneous population. CA attempts to maximize heterogeneity between clusters and, at the same time, maximize homogeneity within clusters, in order to obtain grouping of like observations in terms of pre-treatment characteristics, that are different from other groupings. Here, we perform an agglomerative hierarchical clustering which proceeds sequentially starting from k singleton and composing them in partition tree of nested nodes. In particular, we suggest the use of the Ward's algorithm on factorial coordinates where the proximity between two groups is taken to be the square of the Euclidean distance between them. We favor the Ward's algorithm because, as reported in Lebart et al. (1997), it is based on a optimization criterion (definition of inertia) similar to that on which Multiple Correspondence Analysis is based.

[4] This classification method is also known as Tandem Approach (Arabie and Hubert 1994)

[5] More details about *plus* and *minus* of Tandem Approach could be found in Lebart et al. (1997), Arabie and Hubert (1994)

Table 1 Simulated effects

Combinations	Y(1)	Y(0)	ATE
$x_1 = 1, x_2 = 1, x_3 = 1$ $x_1 = 2, x_2 = 1, x_3 = 1$	$Y(1) = 0.3x_1 + 7x_2 + 3.2x_3$	$Y(0) = Y(1) - 10.62$	10.62
$x_1 = 1, x_2 = 3, x_3 = 2$ $x_1 = 2, x_2 = 3, x_3 = 2$	$Y(1) = 0.88x_2 + 3.33x_3$	$Y(0) = Y(1) + 9.3$	−9.3
$x_1 = 1, x_2 = 2, x_3 = 1$ $x_1 = 2, x_2 = 2, x_3 = 1$	$Y(1) = 6x_1 + 3.3x_2 + 4.1x_3$	$Y(0) = Y(1) - 19.7$	19.7
$x_1 = 1, x_2 = 3, x_3 = 1$ $x_1 = 2, x_2 = 3, x_3 = 1$	$Y(1) = 2.3x_1 + 0.99x_2 + 3x_3$	$Y(0) = Y(1) - 8.45$	8.45
$x_1 = 1, x_2 = 1, x_3 = 2$ $x_1 = 2, x_2 = 1, x_3 = 2$ $x_1 = 1, x_2 = 2, x_3 = 2$ $x_1 = 2, x_2 = 2, x_3 = 2$	$Y(1) = 7x_1 + 0.5x_3$	$Y(0) = Y(1) - 13.97$	13.97

The primary concept of such an approach consists in getting finer cluster partition because it enhances the plausibility of obtaining balanced groups. Once selected the $k-$clusters partition, balance is evaluated and tested within each group. On the basis of test results, local treatment effects will be measured within balanced groups pruning observations in unbalanced clusters.

4 Example

In this section we illustrate with a toy example how the proposed approach works in measuring and testing global balance. We will report results in terms of achieved balance by local spaces. We simulated without error three categorical covariates: X_1 with two levels, X_2 with three levels and X_3 with two levels. We considered a binary assignment-to-treatment indicator variable T(0/1). All possible combinations of covariates[6] have been considered. Units within each of those 12 combinations of covariates were assigned on the basis of different proportions to different treatment levels (t = 0/1), in order to create dependence between **X** and **T**. Since in real applications treatment effects are expected to vary, we simulated heterogeneous treatment effects. To do that, we generated different potential outcomes ($Y(0)$ and $Y(1)$) for different set of covariates combinations (Table 1).

As reported in Table 1 five different average treatment effects (ATE) exist. The *no omitted variable bias* assumption underlying the simulated data assumes a crucial role and thus must be emphasized. By design, the assignment mechanism is assumed to be perfectly known which means that the **X** matrix includes all pre-treatment

[6] $2 \times 3 \times 2 = 12$

Table 2 The local causal effects estimation (k = 1, 2, 3, 4, 5)

Groups	n	$n_{T=1}$	$n_{T=0}$	I_b	Interval for I_b	Balance	ATE
1	355	106	249	0.04	(0;0.008)	no	11.77
2	570	393	177	0.03	(0;0.007)	no	5.87
1	98	47	51	0	(0;0.03)	yes	17.74
2	472	346	126	0.02	(0;0.008)	no	2.85
3	355	106	249	0.04	(0;0.008)	no	11.77
1	240	48	192	0	(0;0.01)	yes	13.97
2	115	58	57	0	(0;0.02)	yes	10.42
3	98	47	51	0	(0;0.03)	yes	17.74
4	472	346	126	0.02	(0;0.008)	no	2.85
1	360	288	72	0	(0;0.008)	yes	8.45
2	112	58	54	0	(0;0.03)	yes	−3.78
3	240	48	192	0	(0;0.01)	yes	13.97
4	115	58	57	0	(0;0.02)	yes	10.99
5	98	47	51	0	(0;0.03)	yes	17.74

variables associated with both the treatment assignment and the observed outcome. We are interested on measuring the Average Treatment Effect. Since, it is well known (Morgan and Winship 2007) that the naive estimator[7] ($\hat{\delta} = 8.2$) is an inconsistent and biased estimate of the ATE, the main aim is to reproduce, as close as possible, the unbiased ATE reported in Table 1. We begin by computing the GI measure for this dataset. Results show that $I_B = 0.15$ and $MIC = 11.26\%$. Under the null hypothesis the interval of plausible values for I_B is $(0; 0.0045)$. The GI measure falls in the critical region and can be interpreted as demonstrating the presence of imbalance in data, thereby demanding adjustment in order to compute unbiased treatment effects. The second step in our analytic process is to use cluster analysis to identify homogeneous groups on the basis of MCA coordinates. The basic idea is that given a k-clusters partition represented in a dendrogram,[8] the lower is its cut level (maximum number of groups), the higher is the possibility of achieving balance. CA was carried out in the SAS system and uses the Ward method and the Euclidean distance as its dissimilarity measure. We most closely examined 12 cluster solutions. We retain $k = 5$ clusters because it provides balance within all clusters (Table 2). Results show that in those clusters the test detects balance suitable unbiased estimates of the local ATE are obtained (Table 2).

Clearly, completely unbiased estimates of local ATE could be obtained by retaining k = 12 clusters (Table 3), where units assigned to a specific covariates combinations fall in the same cluster (Correctly classified rate = 100%).

[7] The naive estimator is defined as the difference in the means of the observed outcome variable for treated and controls in the whole population.

[8] The dendrogram is also called tree-diagram. It consists in a visual representation of a hierarchical clustering procedure.

Table 3 The local causal effects estimation (k = 12)

Groups	n	$n_{T=1}$	$n_{T=0}$	I_b	Interval for I_b	Balance	ATE
1	28	13	15	0	(0;0.11)	yes	13.97
2	50	50	100	0	(0;0.36)	yes	13.97
3	30	30	60	0	(0;0.03)	yes	10.62
4	8	7	15	0	(0;0.17)	yes	13.97
5	14	16	30	0	(0;0.08)	yes	19.7
6	15	15	30	0	(0;0.10)	yes	19.7
7	15	17	32	0	(0;0.11)	yes	8.45
8	17	13	30	0	(0;0.13)	yes	−9.3
9	20	20	40	0	(0;0.06)	yes	10.62
10	26	24	50	0	(0;0.07)	yes	−9.3
11	282	72	360	0	(0;0.008)	yes	8.45
12	48	192	240	0	(0;0.01)	yes	13.97

Working with categorical data the number of possible covariates combinations introduced in the analysis is always known. As a consequence, the k clusters to retain in order to obtain unbiased estimates of local ATE is always known even in real studies. We assert that practical problems related to the number of clusters to retain could arise when the number of covariates used in the analysis is high and a more parsimonious solution in terms of number of clusters could be needed. In such situation we suggest the use of a statistical criterion for select the suited number of clusters combined with an operative criterion based on imbalance test results.

5 Concluding Remarks

The main aim of this paper has been to introduce a Global Imbalance Measure and a Multivariate Imbalance Test. Both those tools were combined within a strategy that helps measuring unbiased treatment effects by subgroups. Such strategy involves first identifying whether imbalance exists, then performing CA on MCA coordinates and finally comparing treatment and comparison cases within balanced clusters to compute treatment effects. We illustrate this strategy with a toy example and learn that where the test detects balance, unbiased treatment effects are reproduced. The main strength of the proposed strategy is that it can simultaneously account for the dependence relationship of any number of covariates. It uses all available information in the **X** matrix without problem of dimensionality even in the presence of categorical covariates. We assume to deal with categorical baseline covariates starting from the consideration that working with categorical covariates is an unavoidable need due to background knowledge, which tends to be qualitative in the social science. Despite working with categorical covariates represents a strength of the approach, future works might explore the continuous case. The practical advantage of the proposed strategy is that the heterogeneity of treatment effects, if present, is taken into account. Computing an overall average treatment effect for

a heterogeneous group may obscure important impacts among subgroups; or overall impacts may be deemed insignificant when they actually are an accumulation of positive and negative effects among various subgroups. For the reason mentioned above, the procedure could be useful for subgroup analysis since it helps discover for whom treatment works best. Although the example is a bit limited, it aims primarily to demonstrate how to apply the strategy to computing treatment effects in instances where data are unbalanced.

References

Arabie, P., & Hubert, L. (1994). Cluster analysis in marketing research. In: R. P. Bagozzi (Ed.), *Advanced methods of marketing research* (pp. 160–189). Oxford: Blackwell.

Benzécri, J. P. (1973). *L'Analyse des Données*. Paris: Dunod.

Camillo, F., & D'Attoma, I. (2010). A new data mining approach to estimate causal effects of policy interventions. *Expert Systems with Applications*, *37*, 171–181.

Daudin, J. J. (1981). Analyse factorielle des dépendances partielles. *Revue de Statistique Appliquée*, *29*(2), 15–29.

D'Attoma, I. and Camillo, F. (2011). A multivariate Strategy to measure and test global imbalance in observational studies. Expert Systems with Applications, 38(4), 3451–3460.

Escofier, B. (1984). Analyse factorielle en reference a un modèle, application a l'analyse de tableaux d'echanges. *Revue de Statistique Apliquée*, *32*(4), 25–36.

Escofier, B. (1988). Analyse des correpondances multiples conditionelle in: Diday(Ed.) Data Analysis and Informatics: International Symposium Proceedings: 5th, Amsterdam: North Holland.

Estadella, J. D. , Aluja, T., & Thiò-Henestrosa, S. (2005). Distribution of the inter and intra inertia in conditional MCA. *Computational Statistics*, *20*, 449–463.

Gibson, C. M. (2003). Privileging the participant: The importance of Subgroup Analysis in social welfare evaluations. *American Journal of Evaluation*, *24*(4), 443–469.

Greenacre, M. J., & Blasius, J. (2006). *Multiple correspondance analysis and related methods*, Boca-Raton, FL: Chapman and Hall.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*, 199–236.

Iacus, S. M., King, G., & Porro, G. (2008). *Matching for causal inference without balance checking*. Available via DIALOG.
http://gking.harvard.edu/files/abs/cem-abs.shtml.

Imai, K., King, G., & Stuart, E. A. (2006). *The balance test fallacy in matching methods for causal inference*. Available via DIALOG.
http://gking.harvard.edu/files/abs/matchse-abs.shtml.

Lebart, L., Morineau, A., & Piron, M. (1997). *Statistique exploratoire multidimensionelle*. Paris: Dunod.

Morgan, S. I., & Winship, C. (2007). *Counterfactual and causal inference: Methods and principles for social research*. University Press, Cambridge.

Peck, L. R. (2005). Using cluster analysis in program evaluation. *Evaluation Review*, *29*(2), 178–196.

Peck, L. R., Camillo, F. and D'Attoma, I. (2010). A promising New Approach to Eliminating Selection Bias. Canadian Journal of Program Evaluation, 24(2), 31–56.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcome Research Methodology*, *2*, 169–188.

Yoshikawa, H., Rosman, E. A., & Hsueh, J. (2001). Variation in teenage mothers' experiences of child care and other components of welfare reform: Selection processes and developmental consequences. *Child Development*, *72*(1), 299–317.

Handling Missing Data in Presence of Categorical Variables: a New Imputation Procedure

Pier Alda Ferrari, Alessandro Barbiero, and Giancarlo Manzi

Abstract In this paper we propose a new method to deal with missingness in categorical data. The new proposal is a forward imputation procedure and is presented in the context of the Nonlinear Principal Component Analysis, used to obtain indicators from a large dataset. However, this procedure can be easily adopted in other contexts, and when other multivariate techniques are used. We discuss the statistical features of our imputation technique in connection with other treatment methods which are popular among Nonlinear Principal Component Analysis users. The performance of our method is then compared to the other methods through a simulation study which involves the application to a real dataset extracted from the Eurobarometer survey. Missing data are created in the original data matrix and then the comparison is performed in terms of how close the Nonlinear Principal Component Analysis outcomes from missing data treatment methods are to the ones obtained from the original data. The new procedure is seen to provide better results than the other methods under the different conditions considered.

1 Introduction

When performing multivariate analysis an *impasse* can be easily reached when dealing with missing data, as the number of dimensions increases and so the number of datapoints. In these cases, statisticians are confronted with the lack of definitive answers in missing data analysis, although the literature on this topic is very rich, starting from the seminal work of Rubin (1976). This is true either in classical or in Bayesian analysis.

The popular solution of simply discarding incomplete cases and carrying out the analysis with the remaining data leads to only a partial knowledge of the phenomenon of interest and it is in contradiction, so to speak, with the modern need of considering all the different facets of information, in order to better speed up the progress of science.

An alternative way of dealing with missing data takes into account the possibility of imputing missing data. Estimated or observed values are used to fill the empty

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_53,

data cells up. The mode imputation or the mean imputation are among these techniques. One can also perform a completely inferential approach which is concerned both with the underlying mechanism that generates the missing data and with the distributional hypotheses assumed for the data. Since this approach is considered more challenging from the theoretical point of view, the imputation approach has gradually lost research interest boost in the last years in favor of the inferential approach.

However, in exploratory multivariate analysis, the inferential approach simply can not be performed. Exploratory techniques that are now so widespreadly used still lack of a fully explored and developed missing value analysis. One of these techniques is the Nonlinear Principal Component Analysis (NLPCA) which can be used, among other possibilities, to extract an overall indicator, or to reduce the dimensionality of data in the case of categorical data.

In this paper we are interested in finding new methods to deal with missing data in explorative multivariate statistical analysis. We focus on the case of NLPCA as a tool to set up an indicator, but our proposal can be easily extended to other multivariate methods. This paper is organized as follows. Section 2 illustrates the essential features of NLPCA, whereas in Sect. 3 a brief description of the main missing data techniques used in NLPCA is presented. Our proposal is introduced in Sect. 4 and tested in Sect. 5, through an *ad hoc* simulation study involving a real situation, that is carried out using the R statistical software. Section 6 summarizes and concludes the paper.

2 The Use of NLPCA to Get Statistical Indicators

The measures of performance, efficacy or efficiency have become of great interest in the statistical research community, since there is an increasing demand of dealing with such issues from many parts of the global society, from policy makers to market operators, from financial experts to healthcare decisors, especially in fields like public opinion polls, in marketing analysis, customer care analysis and so on. So we test our proposal in frameworks where complex indicators have to be extracted from a dataset. NLPCA is often used as a tool to obtain statistical indicators to measure a latent phenomenon which is supposed to lie in categorical data. Consequently we choose to study our proposal when extracting indicators through NLPCA.

NLPCA is a method to measure a latent phenomenon; starting from observed variables, the point of NLPCA is the minimization of a loss function which considers three *ingredients*: the *scores* (values of latent variable), the *category quantifications* for each observed variable and the *loadings* (weights of the observed variables) (see Gifi, 1990 or Michailidis and De Leeuw, 1998 for details).

Formally, let X be such latent variable we want to extract from n units and m ordinal variables with k_j categories ($j = 1, \ldots, m$). Let $\mathbf{G}_j$ be the indicator matrix (of dimension $n \times k_j$) for variable j. The values of X are obtained by minimizing the following quadratic loss function:

$$\sigma(\mathbf{x}, \mathbf{q}_1, \ldots, \mathbf{q}_m) = \frac{1}{m} \sum_{j=1}^{m} (\mathbf{x} - \mathbf{G}_j \beta_j \mathbf{q}_j)'(\mathbf{x} - \mathbf{G}_j \beta_j \mathbf{q}_j), \quad (1)$$

where vector $\mathbf{x}$, of dimension $n \times 1$, contains the *object scores*; vector $\mathbf{q}_j = (q_{j1}, q_{j2}, \ldots, q_{jk_j})'$, of dimension $k_j \times 1$, contains optimal *category quantifications* for variable j; β_j is the *loading* of variable j. Under the following normalization constraints:

$$\mathbf{u}_n'\mathbf{x} = 0; \quad \mathbf{x}'\mathbf{x} = n; \quad \mathbf{u}_{k_j}'\mathbf{D}_j\mathbf{q}_j = 0; \quad \mathbf{q}_j'\mathbf{D}_j\mathbf{q}_j = n, \quad \forall j = 1, \ldots, m \quad (2)$$

where $\mathbf{u}_n$ is a vector of ones of dimension n, $\mathbf{u}_{k_j}$ is a vector of ones of dimension k_j and $\mathbf{D}_j = \mathbf{G}_j'\mathbf{G}_j$ is the $k_j \times k_j$ matrix which contains the frequencies of variable j in its main diagonal, NLPCA produces standardized object scores, standardized quantified variables and a nice interpretation of β_j, that becomes the correlation coefficient between object scores and the quantified variable j. The object scores are the values of the requested indicator.

3 Missing Data Treatment in NLPCA: Existing Methodology and Software Packages

SPSS and R are two of the most popular statistical packages used to deal with missing data in NLPCA. SPSS implements different strategies for missing data. The first (and default) option is just to exclude missing values from the analysis ("passive treatment"). This is possible because the NLPCA solution is not derived from the correlation matrix (which cannot be computed with missing values), but from the data itself. The second option is to impute the mode of the variable (or an extra category) to missing values: this is called "active treatment". This implies that objects with the missing value on this variable are imputed with the same (observed or extra) category. The third option, known as *listwise deletion* method, excludes units with missing data from the analysis (analysis of complete subsets). The R package *homals*, which can perform NLPCA as a particular case of homogeneity analysis, can only implement the *passive treatment*.

4 The Proposed Imputation Procedure

Our proposal (in the following, Forward Imputation) is based upon an iterative algorithm which alternates the performing of NLPCA on a subset of data with no missing values (complete matrix), and the imputation of missing data cells with the corresponding values of the nearest unit in the complete matrix. This sequential process starts from the unit with the lowest number of missing cells and ends with the unit

with the highest number of missing cells. Let $\mathbf{A}$ be an initial data matrix of dimension $n \times m$, affected by missing data. The proposed procedure goes through the following steps:

1. Split matrix $\mathbf{A}$ into a $n_0^{(0)} \times m$-dimensional matrix $\mathbf{A}_0^{(0)}$ with no missing elements and K disjoint $n_k \times m$-dimensional sub-matrices $\mathbf{A}_k$, $k = 1, 2, \ldots, K$ $(K < m)$, where k is the number of missing elements for each row.
2. For each $\mathbf{A}_k$ and each row in $\mathbf{A}_k$, the nearest neighbor imputation missing data method is implemented with the use of loadings and quantifications from the NLPCA performed on the complete matrix $\mathbf{A}_0^{(k-1)}$. Therefore, the missing data are replaced with the values of the closest observation in the complete matrix $\mathbf{A}_0^{(k-1)}$ in terms of the following weighted Euclidean distance:

$$\min_z \; d(u_i^{(k)}; u_z^0) = \min_z \left(\sum_j \beta_j^{(k-1)} |\mathbf{G}_j(i)\mathbf{q}_j^{(k-1)} - \mathbf{G}_j(z)\mathbf{q}_j^{(k-1)}|^2 \right)^{\frac{1}{2}}$$

with $u_z^0 \in \mathbf{A}_0^{(k-1)}$ and j running on all and only those $m - k$ variables which are observed on both objects $u_i^{(k)}$ and u_z^0. Each new complete row is then appended in the complete matrix $\mathbf{A}_0^{(k-1)}$ in order to produce $\mathbf{A}_0^{(k)}$. This procedure is sequentially performed for $k = 1, \ldots, K$ until a NLPCA on the complete matrix $\mathbf{A}_0^{(K)}$ is performed in order to find the final variable loadings $\beta_j^{(K)}$ and category quantifications $\mathbf{q}_j^{(K)}$, $j = 1, \ldots, m$.

The results are not affected by the order in which the objects $u_i^{(k)} \in \mathbf{A}_k$ are completed. In fact, the imputation of each $u_i^{(k)}$ is based only on the matrix $\mathbf{A}_0^{(k-1)}$ and the update to $\mathbf{A}_0^{(k)}$ is carried out when all $u_i^{(k)}$ are completed. This procedure, which is synthetically displayed in Fig. 1, has been implemented in R.

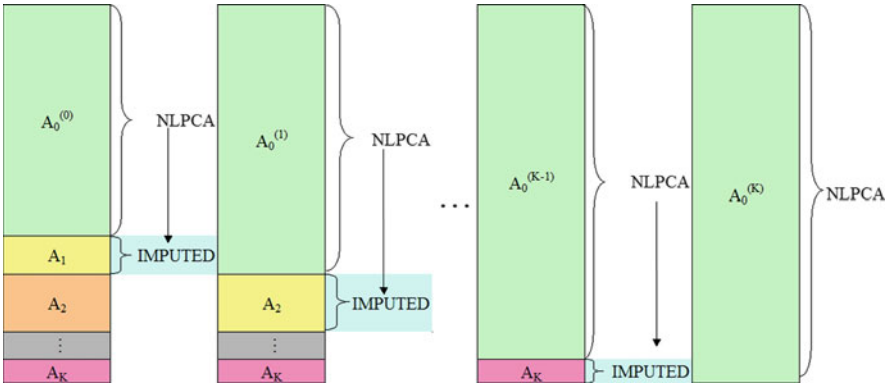


Fig. 1 Graphical representation of the Forward Imputation procedure

Our method presents some important properties. It makes use of the whole information contained in the data; in addition, since it is performed on complete matrices only, it preserves the interpretation of the variable loadings as correlation coefficients and takes into proper account the role of the variable loadings in the minimization problem as a *weighting tool* for the similarity process. Moreover, it includes objects in different steps according to their number of observed variables, ensuring that the role of each object in producing and interpreting results is tied to the number of its actual observations.

On the contrary, the listwise method allows to deal with complete matrices but produces an obvious loss of information. Mode or extra category method, while imputing the same value to all the missing values on the same variable, does not take into account the association with the other variables and reduces the variability on each variable considered. Passive treatment, while preserving the whole information, ignores missing values, works on incomplete matrices, and may lead to an incoherent interpretation of the results.

5 An Application Study with Simulated Missingness Pattern

The relative performance of our proposed method against its main competitors is evaluated through a simulation study, using an *ad hoc* R script.

A specific complete dataset ("true" matrix) is firstly considered; then some observations are removed at random. NLPCA is computed on the "true" dataset and on the dataset with artificial missing data, adopting in the latter case different missing data treatment methods: the Forward imputation (Fo), Passive treatment (Pa), Listwise (Lw) and Mode imputation (Mo), and the outcomes are compared. More specifically the focus is on loadings, which provide the weights of the variables in setting up the indicator, and scores, which allow to classify the units.

5.1 Dataset

The real dataset refers to Eurobarometer 62.1, a survey carried out on behalf of the European Commission, covering the population of the European Union Member States aged 15 years and over. Our attention is focused on Services of General Interest, already analyzed, for example, in Ferrari et al. (2010). Specifically, our analysis refers to the Italian case, year 2004 (905 Italian citizens) and to the following public services: fixed telephone; electricity supply; postal service. We take into account four aspects for each service: service quality, service information, service contracts, customer service, so that 12 variables (four aspects for each of the three services) are eventually collected. The number of possible answers to each question/aspect varies from two to four. The answers are recorded on an ordinal scale.

The goal of NLPCA performed on this dataset is to set up an overall indicator of customers' satisfaction for services of general interest taking into account different services and facets of the service.

The dataset is characterized by a specific missing pattern: all the aspects of the same service are missing at the same time. The percentages of missing units are 28.4% for fixed telephone service, 26.4% for electricity and 28.2% for postal service.

The NLPCA on the complete matrix $\mathbf{Y}$ of 593 Italian citizens on 12 variables produces $\boldsymbol{\beta}$, $\boldsymbol{x}$ and $\boldsymbol{q}_j$ that provide $\rho_{\min} = 0.143$, $\rho_{\max} = 0.699$ and $\lambda_{\max} = 5.083$ (42.3% of the total variance), where ρ represent the pairwise linear correlation coefficient among the quantified variables, and $\lambda_{\max}$ the maximum eigenvalue provided by NLPCA. These values reveal that the analysis gives a good synthesis of the observed variables.

5.2 Simulation Design and Comparison Methods

In order to compare different procedures for handling missing data in different experimental situations, two missing data generating mechanisms are considered:

- Missing Completely At Random (MCAR): 10% (and 20%) of missing data completely at random have been generated on the complete subset $\mathbf{Y}$. For generating the missing data from matrix $\mathbf{Y}$, we draw without replacement a simple random sample of size $v = 713$ (1,426) from the 593×12 matrix, their values are deleted and considered missing.
- Missing by "blocks": the real missingness mechanism that affects the original data is here mimicked, i.e. if one variable (related to a certain service group) is missing, all the variables related to the same service are missing too. With this aim, 28.4% of the 593 units is chosen at random and the values of the four variables of the fixed telephone service are deleted; in the same way, 26.4% of the 593 units is chosen at random and the values of the four variables of the electricity service presented by these units are deleted and 28.2% of the 593 units is chosen at random and the values of the four variables of the postal service are deleted.

For both the scenarios and for each of the $nSim = 1{,}000$ simulation runs, a missing pattern is generated, the loadings of a NLPCA performed with the four different techniques (Fo, Pa, Lw, Mo) are determined as well as the object scores for Fo, Pa and Mo. Of course, Lw is excluded because the object scores for all the 593 units are not determinable by Lw.

Our comparison is based on the vectors of loadings $\boldsymbol{\beta}$ and object scores $\mathbf{x}$. For the loadings, the cosine between the vector of loadings $\boldsymbol{\beta}^{(t)}$ for each missing treatment method and the vector of loadings of the "true" complete matrix $\boldsymbol{\beta}$ is computed:

$$\cos(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}) = \frac{\boldsymbol{\beta}'^{(t)}\boldsymbol{\beta}}{|\boldsymbol{\beta}^{(t)}||\boldsymbol{\beta}|},$$

The closer to one the cosine the better the capability of the missing data treatment method to reproduce the "true" loadings in presence of missing data. To compare the object scores the Root Mean Square Error (RMSE) is used:

$$RMSE^{(t)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i^{(t)} - x_i)^2} \tag{3}$$

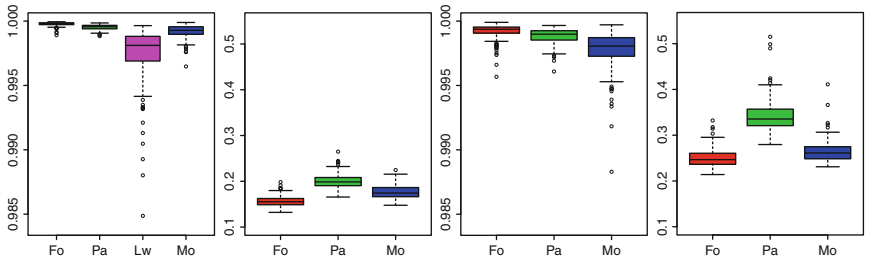
where $x_i^{(t)}$ is the object score for unit i calculated with method t on the matrix with missing data, x_i the object score for unit i calculated on the "true" complete matrix.

The boxplots of both $\cos(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta})$ and $RMSE^{(t)}$ over all the *nSim* simulation runs are provided.

5.3 Results

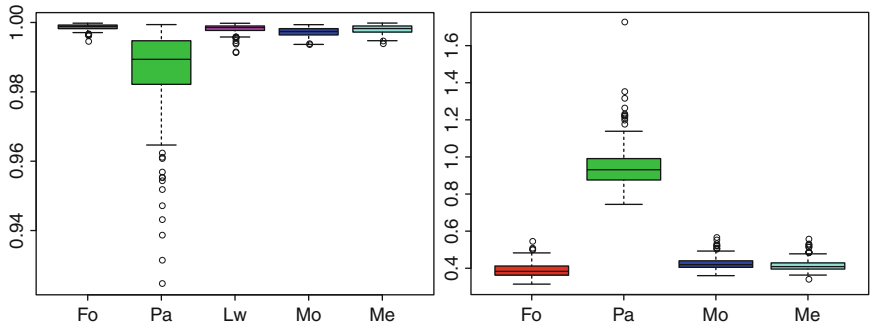
Figure 2 summarizes results based on the performance of the different treatment methods for the first scenario. Specifically, in Fig. 2a, 2b the boxplots of the cosines highlights that the Lw gives the worst performance, while the Fo method always produces a loading vector which is generally the closest to the "true" vector $\boldsymbol{\beta}$. In Fig. 2c the boxplot for Lw is not displayed, since it lies far beneath the other ones. The analysis of the RMSE boxplots in Fig. 2b, 2d confirms the Fo method as the best performer, followed by Mo and Pa.

The results for the second scenario are displayed in Fig. 3 and point out the bad performance of the Pa method, while Fo always performs better than Mo both in terms of loadings and in terms of object scores. The Lw method behaves relatively better in this scenario than in the MCAR setting, but it is worse than our method and in any case provides the scores for classification just for the units with no missing values. The rank of the performance of the methods with regard to RMSE remains the same as in the first scenario.



(a) cos, 10% (b) RMSE, 10% (c) cos, 20% (d) RMSE, 20%

Fig. 2 First scenario (MCAR): comparison of results for different missing data rates (10%, 20%)



(a) Boxplot of cos (b) Boxplot of RMSE

Fig. 3 Second scenario (missing by blocks): comparison of results

6 Discussion and Future Work

In this paper a new proposal for missing data imputation has been described. The proposed method is general, but has been applied here to a multivariate technique (NLPCA) for setting up synthetic numerical indicators. It has been compared to standard techniques for missing data treatment through a simulation study performed on a real dataset.

The simulation study shows that our proposal works better than other missing data treatment methods: it presents a better performance both with MCAR and "missing by blocks" conditions. These results reveal its good performance and its potential generalization and development. In the future we will focus on extending this algorithm to other multivariate methods, on testing it on new missing data conditions and on building up an *ad hoc* R package.

References

Ferrari, P. A., Annoni, P., & Manzi, G. (2010). Evaluation and comparison of European countries: public opinion on services. *Quality & Quantity, 44*(6), 1191–1205.

Gifi, A. (1990). *Nonlinear multivariate analysis*. New York: Wiley.

Michailidis, G., & de Leeuw, J. (1998). The Gifi system of descriptive multivariate analysis. *Statistical Science*, *13*(4), 307–336.

Rubin, D. B. (1976). Inference and missing data. *Biometrica*, *63*, 581–592.

The Brown and Payne Model of Voter Transition Revisited

Antonio Forcina and Giovanni M. Marchetti

Abstract We attempt a critical assessment of the assumptions, in terms of voting behavior, underlying the Goodman (1953) and the Brown and Payne (1986) models of voting transitions. We argue that the first model is only a slightly simpler version of the second which, however, is fitted in a rather inefficient way. We also provide a critical assessment of the approach inspired by King et al. (1999) which has become popular among Sociologists and Political scientists. An application to the 2009 European and local election in the borough of Perugia is discussed.

1 Introduction

Estimating transitions of voters between two adjacent elections is one of extracting information on the association of a two way contingency table from its margins. As shown by Plackett (1977), in the 2×2 case, the information on the odds ratio provided by the margins of a single table is of a rather inconclusive nature. Goodman (1953) provided a formal statistical model which indicates that, under the assumption that a set of local units (which in our context are polling stations) share the same pattern of transitions, this can be consistently estimated from the data. This result relies on the assumption that a set of tables, for which only the margins are observed, are determined by the same probabilistic model.

In the sociological literature the problem is seen as one of inferring individual behaviour from aggregate data and is known as *Ecological Inference*. In a famous paper Robinson (1950) proved that the true underlying association at the individual level and that emerging from aggregate data may even have a different direction, a result known as the Ecological fallacy. As shown in Wakefield (2004, p. 10), this is an instance of the Simpson paradox and may arise when each local unit exhibits an association structure which is strongly correlated with the row marginal. This possibility is ruled out in the Goodman model which, because of its appealing simplicity, is not always applied with sufficient attention to its underlying assumptions.

Brown and Payne (1986) proposed a model which, as we argue below, may be seen as an extension of Goodman's in an attempt to make its assumptions a little

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_54,

more realistic. This model has gained little popularity, perhaps because its estimation procedure is substantially more complex than the linear regression required for Goodman's model. Forcina and Marchetti (1989) reformulated the Brown and Payne model as a multivariate generalized linear model and made available a more efficient software.

An approach popular among Sociologists and Political scientists is due to King et al. (1999, 2004) and is based on a hierarchical Bayesian model. The Bayesian approach proposed by Bernardo (2001) is apparently based on assumptions similar to those of the Brown and Payne model, though the description of the model that he provides is not specified in sufficient detail.

The assumptions underlying Goodman's model and a modified version of the Brown and Payne model are discussed in Sect. 2 and a critical assessment of recent alternative models is given in Sect. 3. An application to transitions between the election for the European Parliament and the one for the borough administration, both held in 2009 is presented in Sect. 4. Concluding remarks are proposed in Sect. 5. The same model has been used to analyze voting transitions for most recent elections held in Umbria (Central Italy), see Bracalente et al. (2006); reports appeared also on the local media.

2 The Goodman Model

Let $\mathbf{n}_u$, $u = 1, \ldots, s$ denote the vector containing the number of voters in local unit u at election 1 ($e1$) and $\mathbf{y}_u$ be the corresponding vector at election 2 ($e2$). Suppose that the voting behavior of voters of party i ($i = 1, \ldots, I$) at $e2$ satisfies the following assumptions:

1. The probability that a voter of party i at $e1$ chooses party j ($j = 1, \ldots, J$) at $e2$ does not depend on the local unit u and is equal to $P(Y = j \mid X = i)$, where X, Y are the options selected at $e1$ and $e2$ respectively
2. Voters decide independently of one another.

Let $\mathbf{y}_{iu}$ denote the vector containing the frequency distribution at $e2$ of voters in unit u who voted party i at $e1$; the above assumptions imply that $\mathbf{y}_{iu}$ is distributed as a multinomial $\text{Mult}(n_{iu}, \mathbf{p}_i)$, where

$$\mathbf{p}_i = (P(Y = 1 \mid X = i), \ldots, P(Y = J \mid X = i))' .$$

The vectors $\mathbf{y}'_{iu}$, $i = 1, \ldots, I$, may be seen as the rows of a frequency table which could be constructed if the choices of each voter at the two elections was known; in reality, only the row and column totals can be observed. However, because $\mathbf{y}_u = \sum_i \mathbf{y}_{iu}$, simple algebra shows that the expectation of the vector of observed proportions $\mathbf{y}_u/(\mathbf{1}'\mathbf{n}_u)$, being the sum of I multinomial random variables, is a linear function of the vectors of transition probabilities $\mathbf{p}_1, \ldots, \mathbf{p}_I$. Thus the transition probabilities could be estimated by multivariate linear regression.

However, the basic assumptions for optimality of ordinary least squares are violated in two directions:

1. The variance of $\mathbf{y}_u/(\mathbf{1}'\mathbf{n}_u)$, the vector of observations in each local unit, equals the variance of a mixture of multinomial variables, thus it is not constant and depends on the unknown transition probabilities
2. Observations within the same local unit are not independent.

Though these violations affect only the efficiency of the estimates, when estimates are adjusted to force values to lie between 0 and 1, consistency of the estimates is also affected.

2.1 The Brown and Payne Model Revisited

Let us consider how realistic are the assumptions on which the multinomial model is based. It seems reasonable to believe that voters may affect each other within small circles; this may be due to personal interactions and to the fact that voters within the same local unit who selected the same party at $e1$ may be affected by common local peculiarities which may be difficult to detect and are naturally treated as random. In both case the multinomial assumption of independence would be violated and a different variance function would be adequate.

The BP model was motivated by the need to take this into account and, at the same time, to avoid the inconveniences due to the method of estimation used to fit the Goodman model. The main features of the model are the following:

1. The vectors of transition probabilities are allowed to differ across local units as in a random effect model, so that now $\mathbf{p}_{i|u}$ denotes the vector of transition probabilities from party i within local unit u; this is formalized by a Dirichelet distribution parameterized with the expectation and the covariance matrix

$$\mathbf{p}_{i|u} \sim \mathit{Dirichelet}(\boldsymbol{\pi}_i, \theta_i[\mathrm{diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i\boldsymbol{\pi}_i'])$$

   in words, transition probabilities fluctuate around the overall average $\boldsymbol{\pi}_i$; direct calculations show that

$$\mathrm{Var}(\mathbf{y}_{iu}) = n_{iu}[1 + \theta_i(n_{iu} - 1)][\mathrm{diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i\boldsymbol{\pi}_i'], \tag{1}$$

   which may be interpreted as the variance of an overdispersed multinomial,
2. Maximum likelihood rather than least square estimates are computed, so the variance structure is taken into account,
3. Transition probabilities $\boldsymbol{\pi}_i$ always lie between 0 and 1 because parameters of interest are defined by a multivariate logit transformation,
4. Because the likelihood for a sum of overdispersed multinomial variables is almost untractable, a central limit approximation is used.

The problem with this approach is that the expression for $\text{Var}(\mathbf{y}_{iu})$, as can be seen from (1), is quadratic in the sample size; this makes the application of the central limit problematic. As an alternative, we propose a model of overdispersion where $\text{Var}(\mathbf{y}_{iu})$ is linear in n_{iu} and where the overdispersion parameter θ_i, as in the Brown and Payne (1986) model, is specific for each party. In the Appendix we show that this variance function is an approximation of the true variance under the following assumptions:

- The voters of party i in local unit u are composed of C_{iu} clusters of size n_{iuc} and their behaviour at $e2$ is determined by a vector of transition probabilities $\mathbf{p}_{iuc}$ which is sampled from a Dirichelet$(\theta_i, \boldsymbol{\pi}_i)$,
- The vector of cluster sizes $\mathbf{m}_{iu} \sim \text{Mult}(n_{iu}, \mathbf{1}/C_{iu})$, that is clusters tend to be of the same size,
- As the n_{iu} increases, n_{iu}/C_{iu} converges to a constant.

3 Recent Alternative Approaches

The hierarchical Bayesian model developed by King et al. (1999) tries to exploit the fact that the frequency distribution of voters at two different elections in a given local unit determines a Frechet class of possible tables of voting transitions consistent with the given margins; it can be shown that, within this class, the transition frequencies may vary within well defined bounds. An advantage of a Bayesian approach is that, because inference is conditional on observations, the estimates of transition within each local unit will always satisfy the bounds implied by the Frechet class. Unfortunately the model is based on a very poor specification of the likelihood as it assumes that, conditionally on transition probabilities specific to each local unit, $\mathbf{y}_u$ is distributed as a multinomial (not a mixture of multinomials); this assumption, which simplifies computations considerably, is equivalent to assume that all voters in local unit u are homogeneous with a common vector of transition probabilities equal to the weighted average of the transitions within each subgroup, a rather unrealistic assumption.

The method proposed by De Sio (2009), which extends to an $I \times J$ table a similar method proposed by Grofman and Merril (2004), is based on King's idea that, if transitions have to be consistent with the observed margins, they are subject to a set of linear constrains. If we require that the estimated transition probabilities satisfy exactly the observed row and column totals on the overall table (obtained by marginalizing over local units) it is likely that the same constraints will be violated in most local units. The method takes as admissible estimates the set of those transition probabilities which are consistent with the observed margins of the overall table and consists in searching for the solution which minimize the sum of squares of violations of the observed constraints across local units. The weakness of this approach seems to be in a lack of a proper statistical model for the random fluctuations which justifies the optimization used in the algorithm.

4 An Application

The revised Brown and Payne model, as described in Sect. 2.1, was applied to estimate transitions between the elections for the European Parliament and the one for the local administration held in the borough of Perugia (PG, central Italy) in June 2009. PG has slightly more than 126 thousands voters and is divided into 159 polling stations; however four of these were removed because they were located inside hospitals or the local prison. Though voters should approximately be the same for the two elections, this is not exactly true as shown by Fig. 1: the two stations with a relative difference larger than 6% were removed. Figure 2 displays the Mahalanobis distance between the results in the local election and those predicted by the model as estimated in a preliminary analysis. The two polling stations with a discrepancy greater than 50 were removed from the final analysis.

Estimated transition probabilities together with standard errors computed from the expected information matrix and a Delta method are displayed in Table 1 where, for conciseness, the original transition matrix based on 10 rows and 12 columns has been condensed.

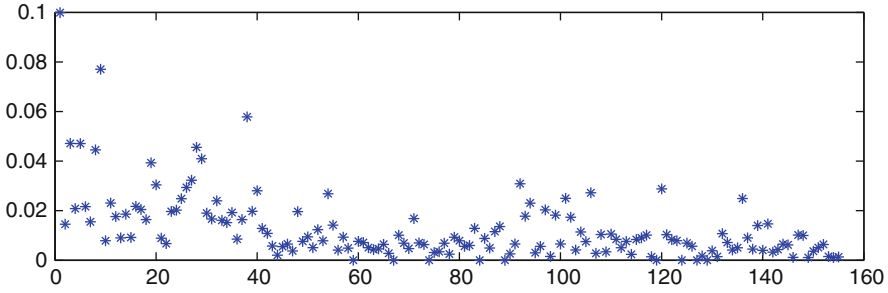


Fig. 1 Relative differences between voters at European and Local election by polling stations

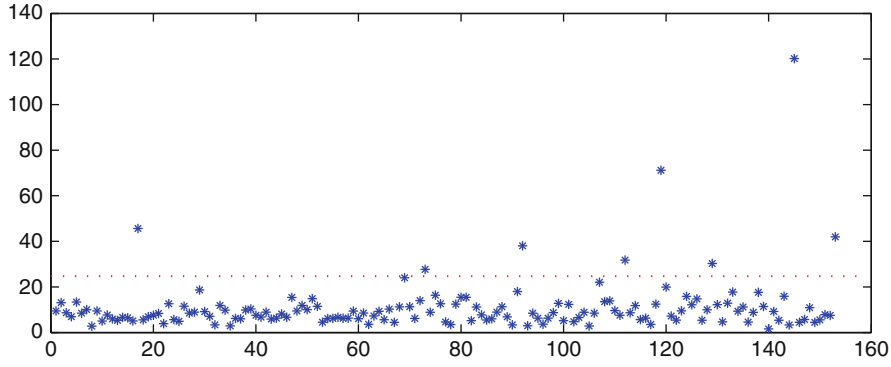


Fig. 2 Mahalanobis distance between observed and predicted electoral results by polling stations and 99% limit

Table 1 Estimated transition out of 1,000 voters and standard errors from the election for the European Parliament (row) to Local administration (column) in Perugia

Party	PD	se	OL	se	LL	se	UDC	se	PdL	se	OR	se	LR	se	NV	se
PD	973	17	9	6	1	1	2	2	11	7	1	1	4	3	0	5
OL	5	10	674	30	146	18	0	15	39	19	19	11	51	22	67	27
UDC	0	1	153	60	0	1	681	52	100	55	0	0	66	42	0	24
PdL	0	5	55	22	14	10	17	11	716	20	5	8	117	18	78	21
OL	4	103	419	101	0	4	0	13	43	77	435	40	98	84	1	63
NV	1	16	0	4	0	0	0	0	0	0	0	1	0	0	999	21

PD = Democratic Party, *OL* = other party on the left, *LL* = local parties on the left, *UDC* = Union of Center Democrats, *PdL* = Party of freedoms, *OR* = other parties on the right, *LR* = local parties on the right, *NV* = non voters

The PD, whose coalition won the local election but in the European election got less votes than the PdL, seems to have a high degree of fidelity. A surprising result is the large proportion of voters who, having supported one of the right wing party in the European election, seems to move to one of the left wing parties in the local election, a phenomenon which, given the local context, is considered plausible; note however that transitions from OR have very large standard errors. Finally note that, though some parties lost voters who abstained or gave a blank ballot, nobody who had abstained in the European election seems to have voted for the local administration.

5 Concluding Remarks

Though the version of the Brown and Payne model presented above could still be improved by considering possible alternative models of overdispersion or by allowing the user to input subjective prior knowledge, it is nevertheless superior to the Goodman's model fitted by linear regression. Relative to the hierarchical Bayesian model of King et al. (1999), our model seems to be based on specific assumptions concerning voting behaviour and does not attempts to provide a general solution to the so called "Ecological Inference". Use of our model is recommended only within areas of limited dimension, like cities of medium size. The reason is that the assumption of a dominating pattern of voting transition would be unrealistic in very large metropolitan areas and even less if applied to a whole country. It is also important that local units, like the Italian polling stations, are reasonably small as the amount of information provided by aggregate data, obviously, decreases when local units of smaller size are clumped together.

In addition, the quality of the data, together with the scope of the application, are a crucial issue. The quality of the data requires that the voters within each local unit are, at least approximately, the same, a requirement which is rather problematic, at least in Italy, because boundaries of local units keep changing from time to time. This could be accommodated simply by merging units which have been redesigned

by internal shifts. Accurate information must also be collected to spot special local units like hospitals whose voters may be expected to be completely different in two different elections and thus must be excluded from the analysis.

It would be desirable if electoral data contained information on the number of new and lost voters between two elections in each unit. However, because such data are not usually available, the most reasonable strategy is to check the absolute relative change in the total number of voters within each unit: if this is below a given threshold and the two elections are close in time, one can simply adjust the data for the second election so that the total number of voters equals that of the first election and remove those units with an absolute relative change above the threshold. This is equivalent to assume that the new voters behave according to a vector of transition probabilities which is a weighted average of the remaining voters, an assumption which can be expected to do little damage as long as the proportion of new voters is small. When estimation is attempted for a large area or a whole country and local units are aggregated within larger administrative boundaries, it will be difficult to check the quality of the data. In addition, it is unlikely that the underlying assumptions are satisfied. As a consequence, the estimated transitions obtained in this way do not provide a consistent estimate of the average transitions for the whole country, even if one had access to accurate data, which is more difficult.

When, like in the Italian system, there is a large number of competing parties and some of them obtain a very small number of votes, two difficulties arise: (i) the normal approximation may not hold when applied to sparse table, (ii) due to the large number of parameters to be estimated, there will be a loss of efficiency and the transitions from small parties will be estimated with large standard errors. In such cases some aggregation of very small parties will be necessary; in any case one should fit a model with as many parties as possible and, if necessary, aggregate and rescale at the end.

Appendix

We give an outline of the proof that, if $\mathbf{y}_{iu} = \sum_c \mathbf{y}_{iuc}$, where $\mathbf{y}_{iuc} \sim \text{Mult}(m_{iuc}, \mathbf{p}_{iuc})$, $\mathbf{m}_{iu} \sim \text{Mult}(n_{iu}, \mathbf{1}/C_{iu})$, $\mathbf{p}_{iuc} \sim \text{Dirichelet}(\theta_i, \boldsymbol{\pi}_i)$ and $\theta_i(n_{iu} - 1)/C_{iu} \to \lambda_i$, then

$$\text{Var}(\mathbf{y}_{iu}) \cong n_{iu}\boldsymbol{\Omega}(\boldsymbol{\pi}_i)(1 + \lambda_i)$$

where $\boldsymbol{\Omega}(\mathbf{x}) = \text{diag}(\mathbf{x}) - \mathbf{x}\mathbf{x}'$ and $\lambda_i \geq 0$.

From (1) $\text{Var}(\mathbf{y}_{iuc} \mid m_{iuc}, \boldsymbol{\pi}_i) = m_{iuc}\boldsymbol{\Omega}(\boldsymbol{\pi}_i)[1 + \theta_i(n_{iuc} - 1)]$; because the $\mathbf{y}_{iuc}$ are independent,

$$\text{Var}(\mathbf{y}_{iu} \mid \mathbf{m}_{iu}, \boldsymbol{\pi}_i) = n_{iu}\boldsymbol{\Omega}(\boldsymbol{\pi}_i)\left[1 + \theta_i(m_{iuc}^2 - 1)/n_{iu}\right].$$

To compute the marginal variance, since $\text{E}(\mathbf{y}_{iu} \mid \mathbf{m}_{iu}) = n_{iu}\boldsymbol{\pi}_i$, $\text{Var}[\text{E}(\mathbf{y}_{iu} \mid \mathbf{m}_{iu})] = \mathbf{0}$ and $\text{Var}(\mathbf{y}_{iu}) = \text{E}[\text{Var}(\mathbf{y}_{iu} \mid \mathbf{m}_{iu})]$ requires only to compute $\text{E}(\mathbf{m}'_{iu}\mathbf{m}_{iu})$;

by a well known result on expectations of quadratic forms, $\mathrm{E}(\mathbf{m}'_{iu}\mathbf{m}_{iu}) = n_{iu}[1 + (n_{iu} - 1)/C_{iu}]$, which gives

$$\mathrm{Var}(\mathbf{y}_{iu}) = n_{iu}\boldsymbol{\Omega}(\boldsymbol{\pi}_i)[1 + \theta_i(n_{iu} - 1)/C_{iu}],$$

the result follows by putting $\lambda_i = \theta_i(n_{iu} - 1)/C_{iu}$.

References

Bernardo, J. M. (2001). *Interpretation of electoral results: A Bayesian analysis*. Internal report, Departamento de Estadistica i I.O., Universitat de Valencia.

Bracalente, B., Ferracuti, L., & Forcina, A. (2006). L'analisi dei flussi elettorali in Umbria: le elezioni dal 2004 al 2006. *AUR&R, 7*, 145–178.

Brown, P. J., & Payne, C. D. (1986). Aggregate data, ecological regression and voting transitions. *Journal of the American Statistical Association*, *81*, 453–460.

De Sio, L. (2009). Oltre il modello di Goodman: l'analisi dei flussi elettorali in base a dati aggregati, *Polena*, Vol. 1, 9–33.

Forcina, A., & Marchetti, G. M. (1989). Modelling transition probabilities in the analysis of aggregate data. In A. Decarli, B. J. Francis, R. Gilchrist, & G. U. H. Seber, (Eds.), *Statistical modelling*. Springer Verlag, Berlin, Heidelberg.

Goodman, L. A. (1953). Ecological regression and the behaviour of individuals. *American Sociological Review*, *18*, 351–367.

Grofman, B., & Merril, S. (2004). Ecological regression and ecological inference. In G. King, O. Rosen, & M. Tanner (Eds.), *Ecological inference*. Cambridge University Press, Cambridge.

King, G., Rosen, O., & Tanner, M. A. (1999). Beta-binomial hierarchical models for ecological inference. *Sociological Methods & Research*, *28*, 61–90.

King, G., Rosen, O., & Tanner, M. A. (Eds.). (2004). *Ecological inference*. Cambridge: Cambridge University Press.

Plackett, R. L. (1977). The marginal totals of a 2×2 table. *Biometrika*, *64*, 37–42.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*, 351–357.

Wakefield, J. (2004). Ecological inference for 2×2 tables. *Journal of the Royal Statistical Society Series A*, *167*, 1–42.

On the Nonlinearity of Homogeneous Ordinal Variables

Maurizio Carpita and Marica Manisera

Abstract The paper aims at evaluating the nonlinearity existing in homogeneous ordinal data with a one-dimensional latent variable, using Linear and NonLinear Principal Components Analysis. The results of a simulation study with Probabilistic and Monte Carlo gauges show that, when variables are linearly related, a source of nonlinearity can affect each single variable, but the nonlinearity of the global solution is negligible and, therefore, can be left out to construct a measure of the latent trait underlying homogeneity data.

1 Introduction

In the social and economic research, the focus is often on the construction of a one-dimensional composite indicator for a latent variable using data from a questionnaire with Likert-type scales. One of the most used procedures to construct the composite indicator is the *summated rating scale*, suggesting to add up the quantifications (usually the first integers) assigned to the ordered response categories of m variables (items) and to use this (weighted or unweighted) sum as a composite measure of a latent construct (Bartholomew 1996). The very simple idea of this procedure is that if the variables are *parallel measurements* (i.e., are *homogeneous* variables) of the same construct, their sum will tend to cancel out measurement errors. Within the variety of procedures aimed at obtaining a one-dimensional indicator from *homogeneous data* (i.e., data from *homogeneous* variables), in the present paper the focus is on Principal Components Analysis (PCA), in its Linear (L-PCA, Jolliffe 2002) and NonLinear (NL-PCA, Gifi 1990; Heiser and Meulman 1994) versions. NL-PCA aims at the same goals of traditional L-PCA, but it is suited for variables of mixed measurement level (nominal, ordinal, numerical) that may not be linearly related to each other. The NL-PCA model is the same linear model as in traditional L-PCA, but it is applied to nonlinearly transformed data, obtained by assigning optimal scale values (the *quantifications*) to the categories. While L-PCA assigns equally spaced numbers (usually the first positive integers) to the categories, NL-PCA finds category quantifications that are optimal in the sense that the overall variance accounted

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_55,

for in the transformed variables, given the number of components, is maximized. In this paper, NL-PCA was applied with the *ordinal scaling level*, meaning that variables are transformed according to monotonic nonlinear transformations. Therefore, in this situation, the difference between the two algorithmic procedures is that L-PCA assigns *linear* (equally spaced) *quantifications*, while NL-PCA assigns *nonlinear* (i.e., not necessarily equally spaced) *quantifications* to the ordered categories; NL-PCA quantifications take also into account possible *nonlinear* relationships among variables (Heiser and Meulman 1994). However, in some applications, NL-PCA leads to the same results of L-PCA, suggesting that the assumptions of L-PCA are not a practical problem (see, e.g., Carpita and Manisera 2006). Starting from these considerations, in this paper our research question is: how much nonlinear are homogeneous data? In order to answer this question we used the *gauging* approach (see Gifi 1990, p. 34): (1) a *Probabilistic gauge* to construct a population of homogenous data and compare the L-PCA and NL-PCA solutions in order to evaluate the level of nonlinearity existing in these data; (2) a *Monte Carlo gauge*, to compare the sampling performance of L-PCA and NL-PCA in recovering the population parameters of interest.

2 Probabilistic and Monte Carlo Gauges

According to Gifi (1990), we used *gauging* to have a realistic data structure with known properties allowing the comparison of the L-PCA and NL-PCA results and the consequent assessment of the nonlinearity. The more L-PCA and NL-PCA results differ, the more nonlinear data are. Population and sampling performances of the two techniques are compared on the category quantifications, i.e., the values assigned to the categories of each variable, and some parameters related to the global solution: the dominant eigenvalues and the *scores*, the composite indicator that can be used as a measure of the latent variable. The method we used to construct population homogeneous data with a one-dimensional latent trait underlying m ordinal variables was proposed by van Rijckevorsel et al. (1985) and is based on the discretization of m continuous variables following a multivariate standard normal distribution with equal correlations ρ. (As in van Rijckevorsel et al. 1985, we chose exactly equal correlations to work with population homogeneous data with a one-dimensional latent trait. This is consistent with the common situation of sample data showing slightly different levels of correlations.) In this case, the m continuous variables are linearly related each other and the correlation matrix has a dominant eigenvalue given by $\lambda_+ = [1 + (m-1)\rho]/m$ (van Rijckevorsel et al. 1985, p. 7). We focused the analysis on the dominant eigenvalue, because other indices used in the classical item analysis depend on its value: for example, the mean correlation $\rho_+ = (\lambda_+)^{1/2}$ and the well-known Cronbach's alpha $\alpha = m(\lambda_+ \cdot m - 1)/[(m-1) \cdot \lambda_+ \cdot m]$ (Heiser and Meulman 1994, p. 187). In this

study we considered[1] $m = 4$ and three different values of $\rho = 0.4, 0.6, 0.8$, corresponding to three situations with underlying one-dimensional latent traits having different levels of strength $\lambda_+ = 0.55, 0.70, 0.85$. Then the continuous variables were discretized, by mapping continuous intervals into ordinal categories using discretization cuts. We considered three discretization procedures resulting from nonlinear monotonic transformations and providing three distributional forms for the ordinal variables: one is the optimal discrete distribution O, which resembles the original normal distribution rather closely; the other two discretization procedures distort the normal distribution resulting in right-skewed discrete distribution (R, with positive skewness) and left-skewed discrete distribution (L, with negative skewness). Following van Rijckevorsel et al. (1985) and considering that few categories allow a higher degree of discretization and therefore of nonlinearity, we chose $k = 5$ categories for each variable with corresponding frequencies (0.11; 0.24; 0.30; 0.24; 0.11) for the O distribution. Two different versions of the skewed variables were considered: *version* (*i*) (0.45; 0.25; 0.15; 0.10; 0.05) and (0.05; 0.10; 0.15; 0.25; 0.45) for the R and the L distributions, respectively, and *version* (*ii*) (0.65; 0.15; 0.10; 0.05; 0.05) and (0.05; 0.05; 0.10; 0.15; 0.65) for the R and the L distributions, respectively. With *version* (*ii*), we chose to stress the presence of high frequencies on the mode category, because quantifications typically show nonlinearity and instability in the presence of low frequencies on some categories, like in the R and L distributions (Linting et al. 2007). Moreover, when the distributions of the analysed ordinal variables are very different, these can be thought to be not linearly related. To evaluate the interaction of ordinal variables with different distributions, we combined optimal and right and left skewed variables, obtaining nine distinct Cases: OOOO, OOOL, OOLL, OLLL, LLLL, LLLR, LLRR, LROO, LLRO. Since the R variable follows the reversed frequency pattern of the L variable, OOOR, OORR and ORRR Cases can be skipped from the analysis, because they give the analogous results of the OOOL, OOLL and OLLL Cases already considered; the same holds for the LRRR and RRRR combinations, equivalent to the LLLR and LLLL Cases, respectively. Using the discretization cuts, we computed the probability distribution for the multinomial distribution by multidimensional integration of the multivariate normal distribution. Then, a population composed of $100,000$ units[2] was obtained for each of the $9 \times 3 = 27$ considered combinations – nine different Cases for O-L-R and three different values of ρ – for the two different *versions* (*i*) and (*ii*) of the skewed R and L distributions, and both L-PCA and NL-PCA were applied to population data.

In order to assess the nonlinearity of the population data, we firstly compared the quantifications obtained by L-PCA (given by the first $k = 5$ positive integers, standardized in order to have zero mean and unit variance) with those given by

[1] We chose this value for m in order to deal with simpler computations but also to introduce some instability in the results, with the aim to stress the differences between L-PCA and NL-PCA.

[2] To simplify and fasten the computation, we chose to compute the parameters of interest from a population data matrix of $100,000$ units rather than by the analytical model, after having checked that that size allowed us to replicate the results reported in van Rijckevorsel et al. (1985).

NL-PCA (assigned by the procedure according to the optimal scaling algorithm). In order to make the comparison between linear and nonlinear quantifications easier, we referred to the *NL* Index (Manisera 2006), which measures the nonlinearity of a given variable using the average squared Euclidean distance between the vectors of the linear $\mathbf{y}_L$ and nonlinear $\mathbf{y}_{NL}$ quantifications assigned to the k categories, weighted by the marginal frequencies:

$$NL^* = \frac{1}{2} \cdot n^{-1}(\mathbf{y}_{NL} - \mathbf{y}_L)'\mathbf{D}(\mathbf{y}_{NL} - \mathbf{y}_L) = 1 - r$$

where n is the number of units, $\mathbf{D}$ is the $k \times k$ diagonal matrix containing the marginal frequencies, and r is the Pearson linear correlation coefficient between $\mathbf{y}_{NL}$ and $\mathbf{y}_L$. Because in the present context both the linear and the nonlinear transformations are non-decreasing, $0 < r \leq 1$ and therefore $0 \leq NL^* < 1$. If c_j for $j = 1, 2, \ldots, k$ are the k categories of the variable, the minimum r and the maximum NL^* are obtained when (A) $c_j = c_1$ or (B) $c_j = c_k$ for $j = 2, \ldots, k-1$. Therefore, the index used in the present paper is given by

$$NL = [NL^*/max(NL^*_A, NL^*_B)] \times 100$$

where NL^*_A and NL^*_B are the NL^* Indices associated to (A) and (B).

Secondly, nonlinearity was studied by comparing the values of the dominant eigenvalues obtained by L-PCA and NL-PCA, on which classical item analysis results like mean correlation and Cronbach's alpha depend, as stated at the beginning of this section. Finally, L-PCA and NL-PCA were compared with reference to the *population scores*, providing the "measure" of the underlying latent trait. For each of the considered combinations, the population scores were obtained by computing the weighted sum of the quantifications assigned by L-PCA or NL-PCA with loadings as weights; one score is determined for each of the unique $k^m = 5^4 = 625$ *population profiles* or response patterns.

In order to check the sample stability of L-PCA and NL-PCA, we derived the *Monte Carlo gauge* from the *Probabilistic gauge* described above: we run $R = 1000$ replications of simple random samples with three different sample sizes $n = 250, 500, 1000$ to study the sampling distributions of the two PCA solutions with reference to dominant eigenvalues and scores.

3 Results

Table 1 displays the *NL* Indices obtained in the *Probabilistic gauge* for the O, L and R variables across the 27×2 considered combinations (see, Sect. 2).

As expected, increasing the skewness of the variables from *version* (*i*) to (*ii*) always increased the nonlinearity of the quantifications. Both O and L variables showed negligible nonlinearity in Cases OOOO and LLLL, when they were not

Table 1 *NL* Indices for the *Probabilistic gauge*

Cases	Variable	*version* (*i*)			*version* (*ii*)		
		$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$
OOOO	O	0.3	0.1	0.0	0.3	0.1	0.0
OOOL	O	0.4	0.6	1.8	0.7	1.7	5.6
	L	2.9	2.6	2.0	7.1	6.7	5.4
OOLL	O	0.8	1.9	4.9	2.1	5.6	14.1
	L	2.1	1.5	0.7	5.2	3.8	2.0
OLLL	O	1.5	3.5	7.9	4.2	10.2	21.2
	L	1.6	0.9	0.2	3.8	2.2	0.7
LLLL	L	1.2	0.5	0.1	2.7	1.2	0.3
LLLR	L	2.3	2.1	3.0	5.4	4.7	4.6
	R	6.7	10.4	16.7	17.2	27.1	40.7
LLRR	L	4.1	5.6	9.4	10.5	13.9	20.9
	R	4.1	5.6	9.4	10.5	13.9	20.9
LROO	L	3.9	5.0	7.2	10.0	12.3	16.5
	R	3.9	5.0	7.2	10.0	12.3	16.5
	O	0.2	0.0	0.3	0.2	0.1	0.1
LLRO	L	3.0	3.3	4.6	7.4	7.8	8.9
	R	5.2	7.7	12.4	13.4	19.5	29.9
	O	0.3	0.4	1.4	0.7	1.5	3.7

combined with variables of other types. Moreover, this nonlinearity decreased as ρ increased and the linear relation among the variables was stronger.

Results referred to the Cases combining variables of two different types showed that the nonlinearity of each variable increased, compared with the Cases combining variables of one single type. Moreover, the nonlinearity of the variable of one type increased when the number of variables of the other type increased: for example, with *version* (*i*) and $\rho = 0.4$, the *NL* Index for variable L was equal to 1.2, 1.6, 2.1, 2.9% in the LLLL, OLLL, OOLL, OOOL Cases, respectively. Looking at the Cases involving only O and L variables, when ρ increased, the value of the *NL* Index decreased for the L variable while increased for the O variable. This result could be explained as follows: when ρ increased, the number of response patterns involving the mode of the skewed variables *and* the highest categories of the O variables increased. In this situation, the highest categories of the optimal variable received NL-PCA quantifications that are close to each other and different from the ones assigned to the lower categories. Consequently, the presence of the L variable implied the nonlinearity of the O variable, in terms of not equally spaced quantifications. The Cases involving combinations of only skewed variables are peculiar: here the *NL* Index reached the highest values in this study. In the LLLR Case with *version* (*ii*) and $\rho = 0.8$, the *NL* Index associated to the R variable was equal to 40.7%; in the LLRR Case, both the R and the L variables showed the same value of 20.9%. The very same result is due to the presence of the same number of R and L variables. This also happened in the LROO Case. In these situations, the L variables received the nonlinear quantifications $\mathbf{y}_{NL}$ while the R variables received

the nonlinear quantifications $-\mathbf{y}_{NL}$, sorted in ascending order, and the *NL* Indices for L and R variables were equal. In the LROO Case combining items of three different types, the presence of two O variables reduced the nonlinearity of the skewed variables: the comparison of the LLRR and LROO Cases showed that in both situations L and R variables had the same nonlinearity, but this nonlinearity was slightly lower in the LROO Case. The ability of the O variables to reduce the nonlinearity of the skewed variables depends on the number of equal skewed variables comprised in the comparing Cases. For example, the LLRO Case showed that the influence of only one O variable was able to reduce the nonlinearity of the L variable, when compared with the LLRR case; on the contrary, the effect of the O variable was not visible when moving from LLRO to LLLR, where the majority of L variables reduced the nonlinearity of the L variables themselves.

The analysis of the dominant eigenvalue in the *Probabilistic gauge* showed that the discretization led to a loss of information: as expected, the dominant eigenvalue in the discrete data was lower than λ_+ in the simulated continuous data. In addition, the study showed that when the Mean NL Index (computed averaging over the different types of variables in each Case) increased, both L-PCA and NL-PCA dominant eigenvalues decreased, indicating that the loss of information was higher in data associated with higher nonlinearity. With respect to L-PCA, NL-PCA was always associated to a lower loss and this best performance of the nonlinear technique directly derives from the NL-PCA optimality criterion. Table 2 displays the percentage variations of the dominant eigenvalue obtained by NL-PCA compared with the one obtained by L-PCA, in correspondence of the 27×2 combinations considered.

As expected, all the variations were positive: the eigenvalue obtained by NL-PCA was always higher than the eigenvalue by L-PCA. This also means that item analysis results like mean correlation and Cronbach's alpha based on NL-PCA performed better. In all the considered Cases, fixed ρ, the best performance of NL-PCA was more evident when moving towards the most skewed variable (*version* (*ii*)). In addition, with respect to L-PCA, the performance of NL-PCA was better as the

Table 2 % variations of NL-PCA vs L-PCA dominant eigenvalues in the *Probabilistic gauge*

Cases	*version* (*i*)			*version* (*ii*)		
	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$
OOOO	0.1	0.1	0.0	0.1	0.1	0.0
OOOL	0.5	0.6	1.0	0.8	1.3	2.3
OOLL	0.7	1.0	1.6	1.3	2.3	4.1
OLLL	0.7	0.9	1.3	1.4	2.1	3.4
LLLL	0.5	0.3	0.1	0.8	0.5	0.1
LLLR	1.5	2.2	3.6	2.4	3.2	3.8
LLRR	1.9	3.1	5.4	3.2	5.0	7.3
LROO	1.0	1.4	2.2	1.6	2.3	3.2
LLRO	1.3	2.0	3.3	2.3	3.4	4.8

nonlinearity increased. When the NL Index moved from 0% to 40.7% (*version* (*ii*) and $\rho = 0.8$), the NL-PCA dominant eigenvalue grew from 0% to 7.3%, with respect to the L-PCA dominant eigenvalue.

With reference to the analysis of population scores, the linear correlation coefficient $\rho_{L;NL}$ computed between the L-PCA and NL-PCA 625 population scores (weighted by their population frequencies) was higher than 0.96 in all the considered configurations. The minimum value 0.96 came out in the LLRR Case with *version* (*ii*) and $\rho = 0.8$. A strong decreasing linear relationship ($R^2 = 0.945$) between $\rho_{L;NL}$ and the Mean NL Index was found. This suggested that as the nonlinearity associated to the variables decreases, L-PCA and NL-PCA substantially provide the same measure of the latent trait.

Results of the *Monte Carlo gauge* suggested that the NL-PCA and L-PCA performances were comparable in terms of accuracy and efficiency in estimating the population parameters. As the sample size increased, bias and variability associated to the estimator decreased, as shown in Table 3 with reference to the estimator of the dominant eigenvalue (in the discrete data) by L-PCA and NL-PCA.

L-PCA and NL-PCA were equally able to estimate population scores, although the *Monte Carlo scores* obtained by NL-PCA showed slightly higher instability. In addition, to compare the *Monte Carlo scores*, we finally computed the linear correlation coefficient between linear and nonlinear scores of all the 625 response patterns (weighted by their population frequencies to consider their sampling probabilities) in each replication for every considered combination. Then we computed the mean correlation, with the associated standard error, over the replications. It was always higher than 0.96 considering the three sample sizes. As expected, it increased as the sample size increased and, for each sample size, showed the lowest values in correspondence of the LLRR Case with *version* (*ii*) and $\rho = 0.8$, like in the *Probabilistic gauge*. Table 4 displays the values of bias and standard error of the correlation coefficient estimator between L-PCA and NL-PCA scores. Results show that such correlation coefficient was well estimated also with the smallest sample size, and this confirms that L-PCA and NL-PCA were comparable in their ability to estimate the measure of the latent variable underlying the data.

Table 3 Bias and standard error of L-PCA and NL-PCA dominant eigenvalue estimator for the *Monte Carlo gauge* (1,000 replications)

	n	L-PCA			NL-PCA		
		250	500	1000	250	500	1000
Bias	min	−0.002	−0.001	−0.001	0.002	0.001	0.000
	mean	0.000	0.000	0.000	0.008	0.004	0.002
	max	0.003	0.002	0.001	0.017	0.011	0.004
Std Err	min	0.014	0.010	0.007	0.016	0.011	0.007
	mean	0.022	0.015	0.011	0.023	0.016	0.011
	max	0.032	0.023	0.016	0.031	0.022	0.015

Table 4 Bias and standard error of the correlation coefficient estimator between L-PCA and NL-PCA scores for the *Monte Carlo gauge* (1,000 replications)

	n	250	500	1000
Bias	min	0.001	0.000	0.000
	mean	0.006	0.003	0.002
	max	0.015	0.008	0.005
Std Err	min	0.002	0.002	0.001
	mean	0.006	0.003	0.002
	max	0.010	0.007	0.005

4 Concluding Remarks

The results of this study showed that when population data are homogeneous and variables of very different distributional forms are combined, a source of nonlinearity can appear when the attention is on each single variable. However, when the focus is on the global solution and, in particular, on the construction of a measure of the latent trait underlying homogeneous data, nonlinearity is negligible even when variables have very different distributions. For this purpose, L-PCA and NL-PCA results (recovery of parameters and sampling performances) do not practically differ: according to the Occam's razor, the use of the simpler method can be preferred, although ordinal data from variables having different (optimal and skewed) distributions would suggest to use the more complex nonlinear technique. Future research will extend the study to cover more distribution structures, with variables nonlinearly related each other, discretization types and sample sizes, in order to get more information on the issue of nonlinearity.

References

Bartholomew, D. J. (1996). *The statistical approach to social measurement*. London: Academic Press.

Carpita, M., & Manisera, M. (2006). Un'analisi delle relazioni tra equità, motivazione e soddisfazione per il lavoro. In M. Carpita, L. D'Ambra, M. Vichi, & G. Vittadini (Eds.), *Valutare la qualità. I servizi di pubblica utilità alla persona* (pp. 311–350). Milano: Guerini.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester: John Wiley.

Heiser, W. J., & Meulman, J. J. (1994). Homogeneity analysis: exploring the distribution of variables and their nonlinear relationships. In M. Greenacre, & Blasius, J (Eds.), *Correspondence analysis in the social sciences* (pp. 179–209). New York: Academic Press.

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). New York: Springer.

Linting, M., Meulman, J. J., Groenen, P. J. F., & Van der Kooij, A. (2007). Stability of nonlinear principal components analysis: An empirical study using the balanced bootstrap. *Psychological Methods*, *12*, 359–379.

Manisera, M. (2006). Measuring nonlinearity in data analysis. *Proceedings of the XLIII Meeting of the Italian Statistical Society*, pp. 689–692, Padova: CLEUP.

van Rijckevorsel, J., Bettonvil, B., & de Leeuw, J. (1985). *Recovery and stability in nonlinear PCA*. Department of Data Theory, Leiden Univ, RR–85–21, Leiden (The Netherlands).

New Developments in Ordinal Non Symmetrical Correspondence Analysis

Biagio Simonetti, Luigi D'Ambra, and Pietro Amenta

Abstract For the study of association in two and three-way contingency tables the literature offers a large number of techniques that can be considered. When there is an asymmetric dependence structure between the variables, the Goodman-Kruskal and Marcotorchino index (with respect to the Pearson chi-squared statistic) can be used to measure the strength of their association when they are collected in two and three way contingency tables, respectively. In the last years, special attention has been paid to the graphical representation of the dependence structure between two or more variables, preserving the information arising from the ordinal structure of the modalities. In this paper, the authors synthesize the main proposals falling within the framework called Ordinal Non Symmetrical Correspondence Analysis for two and three way contingency tables.

1 Introduction

For the analysis of contingency tables, the Pearson chi-squared statistic is the most common tool used to measure the association between two or more variables. However, in situations where there is a one-way, or asymmetric, relationship between categorical variables it is not appropriate to use the Pearson chi-squared statistic. Instead, for such two way tables the Goodman-Kruskal tau index (Goodman and Kruskal 1954) and one of the multivariate extensions, the Marcotorchino index can be considered. We highlight that when the contingency tables consists of ordinal categorical variables their structure often needs to be preserved. In order to graphically summarize the result arising from the dimensional reduction that is used to the asymmetrical relationship between three ordinal variables and to take into account the ordinal nature of the variables, in the last year has been proposed a class of techniques named Ordinal Non Symmetrical Correspondence analysis (ONSCA, D'Ambra and Lauro, 1989) that provide a graphical description of the dependence structure of the predictor categories on the response categories. Such a summary is also shown to be of benefit when determining how individual categories differ in terms of their location, dispersion and higher order moments.

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_56,

2 Ordinal Correspondence Analysis

One popular method of determining the structure of the association among categorical variables is correspondence analysis (CA), 'an exploratory multivariate technique that converts a matrix of non-negative data into a particular type of graphical display in which the rows and columns of the matrix are depicted as points' (Greenacre and Hastie 1987). The classical approach considers the case where variables are symmetrically related. One of the few limitations of classical CA is that it does not allow for a researcher to consider the case where two categorical variables are asymmetrically associated. For example, the age of a person may influence how they respond to a particular product, but the product will not influence the persons age. In many practical situations the variables will often be related in this manner. When there are only two categorical variables, this problem can be overcome by considering non symmetrical correspondence analysis (NSCA), a variation of classical CA approach. The primary difference between the two procedures lies in the measure of association that is considered-CA involves decomposing the Pearson chi-squared statistic while NSCA decomposes the Goodman-Kruskal tau for two-way table index (Goodman and Kruskal 1954) and Marcotorchino index (Marcotorchino 1984) for three-way contingency table. In the last decades Non Symmetrical Correspondence Analysis (NSCA) has been accepted as a useful tool for graphically describing the relationship among categorical variables with a one-way association. Much of the discussion has focused on its application to variables with a nominal structure. However recent theory developed for classical correspondence analysis has been shown to be applicable in cases where the response variable and predictor variable of a two or three way contingency table are ordinal. Such approach involves a decomposition bases on orthogonal polynomials for contingency table (Emerson 1968) and generalized correlations rather than singular value decomposition. The cited decomposition has generated a class of techniques falling within the framework called Ordinal Non Symmetrical Correspondence Analysis for two and three way data tables highlighting the descriptive and inferential main properties (Table 1).

Table 1 Ordinal non symmetrical correspondence analysis techniques

Technique	Data table	Row	Column	Tube	Index	Decomposition method
2w SONSCA	Two-Way	Ordinal	Nominal	—	Goodman-Kruskal	Hybrid[a]
2w DONSCA	Two-Way	Ordinal	Ordinal	—	Goodman-Kruskal	Polynomials
3w SONSCA	Three-Way	Ordinal	Nominal	Nominal	Marcotorchino	Hybrid[a]
3w DONSCA	Three-Way	Ordinal	Ordinal	Nominal	Marcotorchino	Hybrid[a]
3w FONSCA	Three-Way	Ordinal	Ordinal	Ordinal	Marcotorchino	Polynomials
2w SROCA	Two-Way	Nominal	Nominal	—	Chi-Squared	Hybrid[a]
2w DROCA	Two-Way	Nominal	Nominal	—	Chi-Squared	Polynomials

[a]Mixed approach with SVD and Polynomials. 2w (3w) stands for two (three)-way

2.1 Two-Way Data Table

For the analysis of two-way data table with one or two ordinal variables, Lombardo et al. (2007) propose the decomposition of the Goodman and Kruskal index and the graphical visualization of the data association structure for a two-way data matrix. The methodology presented and referred as single or doubly ordinal non symmetrical correspondence analysis in the case when the data table consists of one or two ordinal variables, respectively, is designed to allow the user to visualize the dependence relationship between categories of a response and a predictor variable. Such a visualization is useful when identifying the structure of this relationship and does so in terms of components that reflect sources of variation in terms of the location (mean), dispersion (spread) and higher order moments.

2.2 Three-Way Data Table

For the analysis of three-way contingency tables with one dependent and two predictors variables, the Marcotorchino index represents an appropriate statistical tool. In addition, for contingency tables that consist of ordinal categorical variables, their ordered structure often needs to be preserved. In this case, Simonetti (2003), Beh et al. (2007) focus on the partition of the Marcotorchino index, identifying sources of variation within each variable in terms of location, dispersion and higher order moments. The authors discuss the case when all three variables are ordinal (completely ordered), or when only one ore two are ordinal and the other nominal (partially ordered case). For the decomposition the authors use an approach based on the decomposition through orthogonal polynomials (completely ordered case) or a mixed (hybrid) approach using orthogonal polynomials and singular value decomposition (partially ordered case).

To graphically describe the relationship between the row (response) variable and the column (explanatory) variable, one may consider the plot of the decomposition of the Marcotorchino index using orthogonal polynomials decomposition as proposed by D'Ambra et al. (2006).

2.3 Constrained ONSCA

As highlighted by several authors (Nishisato 1980; Böckenholt and Böckenholt 1990), introducing linear constraints on the row and column coordinates of a correspondence analysis representation may be greatly simplify the interpretation of the data matrix. The Ordinal CA technique has been shown to be a more useful and informative correspondence analysis method than the classical technique commonly used, even if he restricted the analysis to the integer valued scores. The problem with such a scoring scheme is that it assumes that the ordered categories

are equally spaced. In general we know that this may not be the case. Moreover, in order to obtain a linear order for the standard scores, Böckenholt and Böckenholt (1990) remove the effects of the quadratic and cubic trend in order to obtain a linear order for the standard scores, by including suitable constraint matrices even if they do not know they are statistically significant sources of variation. Main aim of the Restricted ONSCA (Amenta et al. 2008) has been to introduce the constraints to the modalities in order to take into account unequally spaced categories, and to consider linear constraints directly on suitable matrices that reflect only the most important components. Starting from such aim, Amenta et al. (2008) introduce linear constraints to the Beh's Ordinal Correspondence Analysis in the case when the contingency table consists of one ordinal and one nominal variables.

3 Example of Three-way FONSCA: Application on cheese data

We perform a survey on a panel of 125 consumers to evaluate how the appearance and the smell of an Italian cheese influence the overall satisfaction. The survey was conducted using a four point scale for the three variables. The three variables are 'Satisfaction with cheese', 'Appearance of cheese' and the 'Strength of Smell of Fresh Milk' in the cheese. For this data it is assumed that the satisfaction variable is dependent on the appearance of the cheese and smell of the milk. Therefore we will treat the data as asymmetric with the row variable as the predictor and the other two variables as explanatory (predictor). A preliminary analysis of the association between the predictor and explanatory variables, can be made drawing the following Z-plots (Choulakian and Allard 1998) where $Z(i|k)$ are the marginal empirical distribution function values of the ordinal response variable.

Figure 1a shows the relationship between the appearance and the overall satisfaction, a good evaluation for the appearance tend to have an acceptable overall satisfaction. In fact it appears there are very little difference between the Satisfaction levels 3 and 4 (excellent). However it is unclear how distinctive the appearance levels are. Figure 1b shows a strong smell of fresh milk in cheese tends to lead to an excellent satisfaction level. However it is unclear whether there is any difference in the 'smell' categories. To measure the level of dependence in the asymmetric data we compute the Marcotorchino index $= 0.026154$. To determine how all three variables are related to one another the Marcotorchino index can be decomposed into location, dispersion and 'error' components for each variable. This is summarized in Table 2.

When observing the influence of the one explanatory variable on the association between the predictor and second variable observe the values in Table 2. It shows that the location effect of the 'smell' variable account for 83.3% of the association between the satisfaction and appearance variables. Similarly the location effect of the appearance categories is the major influence of the association of the satisfaction and smell. Graphical summaries of the three-way associations can also be obtained.

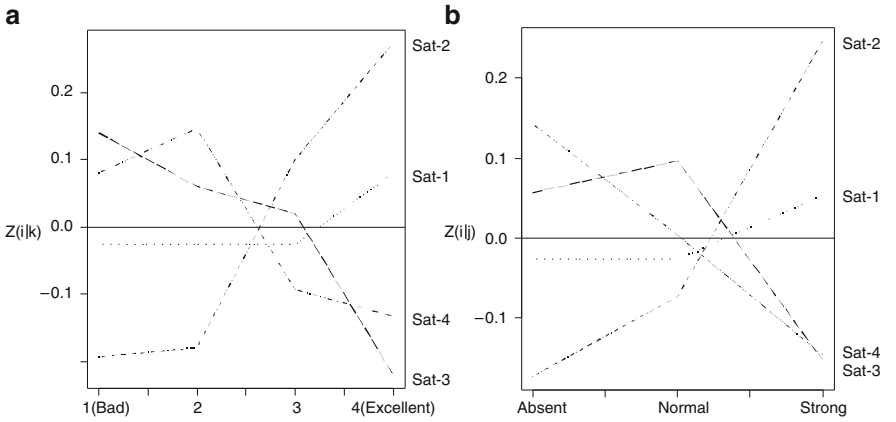


Fig. 1 (**a**) Doubly ordered symmetric correspondence plot of the two predictor variables. (**b**) Joint representation of the three variables

Table 2 Decomposition of Marcotorchino index

Component	Value	Contribution
Row-Column		
Location	0.021787	83.30
Dispersion	0.004367	16.70
Error	0.000000	0.00
Row-Tube		
Location	0.017866	68.31
Dispersion	0.006245	23.88
Error	0.002043	7.81
Marcotorchino	0.026154	100.00

Figure 2a confirms that the smell and appearance variables are positively influential in the satisfaction level of the cheese. That is a cheese with an excellent appearance and has a strong smell of fresh milk can lead to excellent satisfaction levels. Similarly cheeses with a poor appearance and where the smell of fresh milk is absent tend to produce poor satisfaction levels. Figure 2b graphically shows the influence of the smell variable on the level of satisfaction. Since the coordinate along the first axis of the extremes of the satisfaction levels (sad and excellent), they are similarly located. However the difference along the second axis indicates variation in terms of their dispersion. Again Satisfaction levels 2 and 3 appear to have similar classification when the location effect of the appearance is taken into consideration. Suppose we now consider the impact of the location effect of the 'smell' categories on the satisfaction and appearance variables. In the case of classical correspondence analysis, Lebart et al. (1984), presented the idea of confidence circles to identify for which categories the hypothesis of independence is rejected. Beh and D'Ambra (2009) proposed confidence interval for NSCA (see Fig. 3).

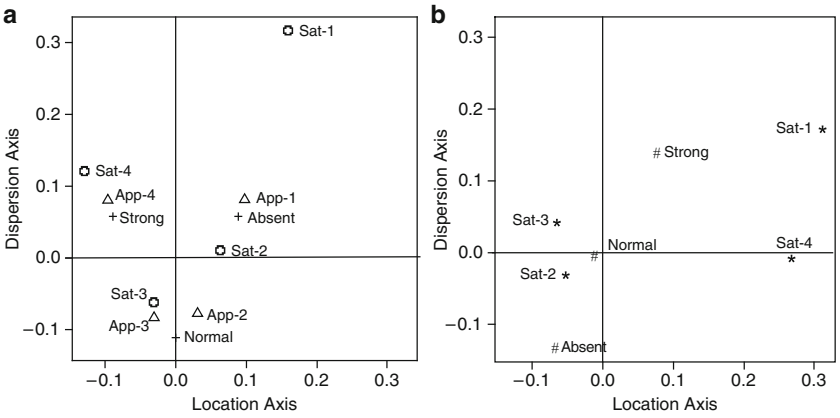


Fig. 2 Graphical summary of predictor and two explanatory variables of Table 1 [Fig. 2a] and Ordinal Non-symmetrical correspondence plot of the satisfaction and smell variables taking into account the points position on the location axis between the appearance categories [Fig. 2b]

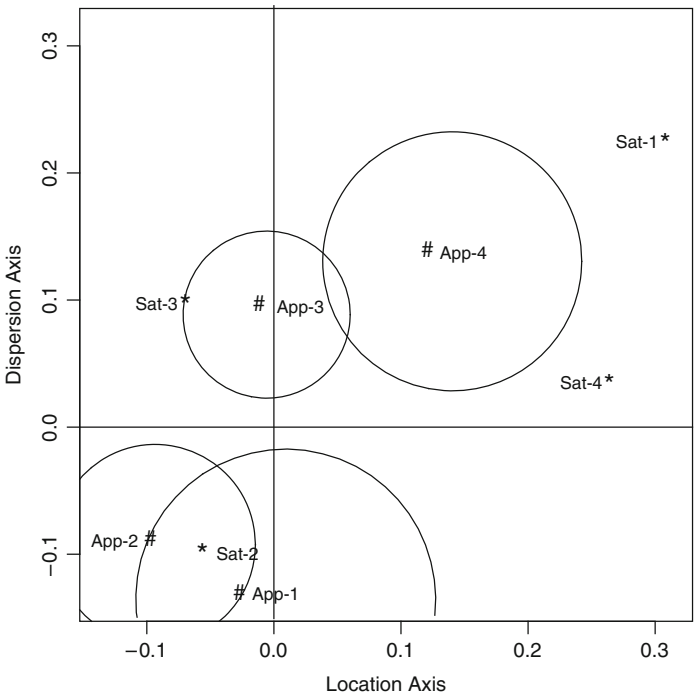


Fig. 3 Ordinal non-symmetrical correspondence plot of the satisfaction and appearance variables taking into account the location effect between the smell categories with 95% confidence regions for the appearance modalities

Therefore, the 95% confidence circle for the *j*th column coordinate in the two-dimensional non-symmetrical correspondence plot has a radius length

$$r_j^J = \sqrt{\frac{5.99\left(1 - \sum_{i=1}^{I} p_{i\bullet}^2\right)}{p_{\bullet j}\,(n-1)\,(I-1)}} \tag{1}$$

Note that (1) depends on the *j*th marginal proportion. Thus, for a very small classification in the *j*th predictor category, the radius length will be relatively large. Similarly, for a relatively large classification, the radius length will be relatively small. We draw on the non-symmetrical plot, the 95% confidence circle for the j^{th} column coordinate. The confidence circles show that all modalities of appearance variable are significant with an error level of 5%.

References

Amenta, P., Simonetti, B., & Beh, E. J. (2008). Single ordinal correspondence analysis with external information. *Asian Journal of Mathematics & Statistics* , *I*(1), 34–42.

Beh, E. J., & D'Ambra, L. (2009). Some interpretative tools for non-symmetrical correspondence analysis. *Journal of Classification*, *26*, 55–76.

Beh, E. J., Simonetti, B., & D'Ambra, L. (2007). Partitioning a non-symmetric measure of association for three-way contingency tables. *Journal of Multivariate Analysis*, *98*, 1391–1411.

Böckenholt, U., & Böckenholt, I. (1990). Canonical analysis of contingency tables with linear constraints. *Psychometrika*, *55*, 633–639.

Choulakian, V., & Allard, J. (1998), The Z-plot: A graphical procedure for contingency tables with an ordered response variable. In J. Blasius & M. Greenacre (Eds.), *Visualization of categorical data* (pp. 99–105). London: Academic.

D'Ambra, L., & Lauro, N. C. (1989). Non-symmetrical correspondence analysis for three-way contingency table. In R. Coppi & S. Bolasco (Ed.), *Multiway data analysis* (pp. 301–315). Amsterdam: Elsevier.

D'Ambra L., Simonetti B., & Beh, E. J. (2006). A dimensional reduction method for ordinal three-way contingency tables. In A. Rizzi & M. Vichi (Eds.), *Proceedings of compstat* (pp. 271–283). Rome: Physica-Verlag HD.

Goodman, L., & Kruskal, W. (1954). Measures of association for cross-classifications. *Journal of the Acoustical Society of America*, *49*, 732–764.

Emerson, P. L. (1968). Numerical construction of orthogonal polynomials from a general recurrence formula. *Biometrics*, *24*, 696–701.

Greenacre, M., & Hastie, T. (1987). The Geometric Interpretation of Correspondence Analysis. *Journal of the Acoustical Society of America*, *82*(398), 437–447.

Lebart, L., Morineau, A., & Warwick, K. M. (1984). *Multivariate descriptive statistical analysis*. New York: Wiley.

Lombardo R., Beh E. J., & DŠAmbra L. (2007). Non-symmetric correspondence analysis with ordinal variables using orthogonal polynomials. *Computational Statistics & Data Analysis*, *52*(1), 566–577.

Marcotorchino, F. (1984). *Utilisation des comparaisons par paires en statistique des contingencies. Partie I*. Report # F 069. France: IBM.

Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. Toronto: University of Toronto Press.

Simonetti, B. (2003). *Ordinal and non ordinal non-symmetric correspondence analysis for three-way tables in sensorial analysis*. PhD Thesis. Napoli: University of Naples Federico II.

Correspondence Analysis of Surveys with Multiple Response Questions

Amaya Zárraga and Beatriz Goitisolo

Abstract Correspondence Analysis (CA) of surveys studies the relationships between several categorical variables defined with respect to a certain population. However, one of the main sources of information is the type of survey in which it is usual to find multiple response questions and/or conditioned questions that do not need to be answered by the whole population. In these cases, the data coded as 0 (category of no chosen response) and 1 (category of chosen response) can be expressed by means of an incomplete disjunctive table (IDT). The direct application of standard CA to this type of table could lead to inappropriate results. We therefore propose a new methodology for the analysis of incomplete disjunctive tables.

1 Introduction

The aim of this work is to present a new methodology to describe simultaneously the relationships between all the categorical variables that make up a survey that contains multiple response questions and/or conditioned questions and that, therefore, give rise to an incomplete data table. Factorial studies of information from surveys are usually carried out by means of multiple correspondence analysis (MCA) (Lebart et al. 1984; Escofier and Pagès 1998; Greenacre 1984). MCA is widely used for analysing surveys with multiple choice questions that contain a finite number of response categories from which individuals have to choose one and only one. This method consists of applying CA to what it is known as a complete disjunctive table (CDT) in which the answers given by individuals are coded as 1 (category of chosen response) and 0 (category of no chosen response). However, apart from multiple choice questions, surveys may often contain multiple response questions in which individuals can simultaneously select more than one category of response. Moreover, there are also surveys in which it is usual to find conditioned questions that do not need to be answered by the whole population. In this last case, individuals must answer a question or not depending on their answer to a previous one. In surveys it is also possible to find a combination of both types of questions, that is to say, multiple response questions conditional upon previous responses. In these cases, information is gathered in a table that is known as an Incomplete Disjunctive Table (IDT).

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_57,

Table 1 Example of an IDT

	A1	A2	B1	B2	B3	C1	C2	D1	D2	D3	D4	
1	1	0	1	0	0	1	0	1	0	1	0	5
2	1	0	1	0	0	1	0	1	0	0	0	4
3	0	1	0	1	0	0	1	0	0	0	0	3
4	0	1	0	0	1	1	0	1	1	1	1	7
...												
	6	4	4	3	3	7	3	7	3	6	4	50

For example, assume that Table 1 presents the answers of 10 individuals to a survey comprising questions A, B, C and D with their corresponding response categories. Note that answers to question D are conditioned by question C. In particular, only individuals who choose category C1 answer question D. Also note that question D is a multiple response question. Consequently Table 1 is an IDT. In this small example the aim would be to describe the relationships between four categorical variables from the answers provided by 10 individuals. This aim can be approached in three different ways, by classical CA of the IDT, by classical CA of the CDT completing the IDT with dummy categories and by the methodology proposed in this paper, which consists of applying CA to the IDT, imposing a suitable marginal.

The notation used is introduced in Sect. 2, and then Sect. 3 presents the problems that result from the application of the first analysis. Section 4 shows the problems that result from the application of the second analysis. Sections 5 and 6 describe the methodology proposed and give an illustrative example.

2 Notation

Let $\mathscr{I} = \{1, \ldots, i, \ldots, n\}$ be the set of individuals, $\mathscr{Q} = \{1, \ldots, q, \ldots, Q\}$ the set of variables (questions) which individuals must answer, $\mathscr{J}_q = \{1, \ldots, j, \ldots, J_q\}$ the set of categories of variable q and $\mathscr{J} = \{1, \ldots, j, \ldots, J\}$ the set of categories of all variables. Let $\mathbf{Z}$ and $\mathbf{Z}^*$ be the complete disjunctive table (CDT) and incomplete disjunctive table (IDT) respectively, both with n rows and J columns, whose general terms are: $z_{ij} = 1$ if individual i answers category j and $z_{ij} = 0$, otherwise. Define $z_{i.} = \sum_j z_{ij}$ as the total number of answers of any individual i, $z_{.j} = \sum_i z_{ij}$ as the number of individuals responding in category j, and finally $z = \sum_i \sum_j z_{ij}$ as the grand total of the table. Note that in $\mathbf{Z}$ there is always just one 1 value and $(J_q - 1)$ zeros for each question q and for each individual i. However, in $\mathbf{Z}^*$ for some questions q, there are J_q zeros for those individuals not answering question q in the case of conditioned questions, and in the case of q being a multiple response question an individual may have from J_q ones to J_q zeros. As a result, some properties given in the CDT are no longer fulfilled in the IDT. Thus, in the IDT the number of individuals who answer each question is not n for all questions, the total number of answers of any individual is not Q and the total of the table is different from nQ.

Relatives and marginals frequencies are defined in the usual way: $p_{ij} = z_{ij}/z$, $p_{i.} = \sum_j p_{ij} = z_{i.}/z$, $p_{.j} = \sum_i p_{ij} = z_{.j}/z$, as are the row profiles: $p_{ij}/p_{i.} = z_{ij}/z_{i.}$, which set up the row-point cloud (individuals) in R^J whose centre of gravity or average row profile coordinate is $p_{.j}$. This centre of gravity or average row profile represents the origin of the factorial axes over which the row-point cloud is projected. Column profiles are defined as $p_{ij}/p_{.j} = z_{ij}/z_{.j}$, which set up the column-point cloud (categories) in R^n whose centre of gravity or average column profile coordinate is $p_{i.}$. This centre of gravity or average column profile represents the origin of the factorial axes over which the column-point cloud is projected.

3 Problems of Applying Classical CA to an IDT

In CA, similarity between any pair of row profiles (individuals) and between any pair of column profiles (categories) is calculated by means of the χ^2 distance.

The χ^2 distance between two row profiles i and i' is defined by:

$$d^2(i,i') = \sum_{j\in\mathscr{J}} \frac{1}{p_{.j}}\left(\frac{p_{ij}}{p_{i.}} - \frac{p_{i'j}}{p_{i'.}}\right)^2 = \sum_{j\in\mathscr{J}} \frac{z}{z_{.j}}\left(\frac{z_{ij}}{z_{i.}} - \frac{z_{i'j}}{z_{i'.}}\right)^2$$

In a CDT two individuals i and i' are similar if overall they have the same response categories. This distance only increases with the different answers (as is logical). But in an IDT this distance also increases with common answers when the individuals do not answer the same number of questions, since $z_{i.} \neq z_{i'.}$.

For example, in Table 1 the response patterns of individuals 1 and 2 differ only in category D3 and this single difference would have to be the only one determining the distance between both individuals. Nevertheless, the application of the above expression to these individuals implies that the distance also increases with all the common answers: $d^2(i=1,i'=2) = \frac{50}{6}\left(\frac{1}{5} - \frac{1}{4}\right)^2 + \cdots + \frac{50}{4}\left(\frac{0}{5} - \frac{0}{4}\right)^2$.

Similarly, the χ^2 distance between two categories j and j' is defined as:

$$d^2(j,j') = \sum_{i\in\mathscr{I}} \frac{1}{p_{i.}}\left(\frac{p_{ij}}{p_{.j}} - \frac{p_{ij'}}{p_{.j'}}\right)^2 = \sum_{i\in\mathscr{I}} \frac{z}{z_{i.}}\left(\frac{z_{ij}}{z_{.j}} - \frac{z_{ij'}}{z_{.j'}}\right)^2$$

In a CDT the marginal on $\mathscr{I}$ is constant, $p_{i.} = 1/n$, so each individual takes part with the same importance in the formation of this distance. But in an IDT each individual could have a different importance according to the number of answers previously chosen, since $p_{i.} \neq 1/n$.

For example, if the above distance between categories D1 and D4 in Table 1 is calculated: $d^2(D1,D4) = \frac{50}{5}\left(\frac{1}{7} - \frac{0}{4}\right)^2 + \frac{50}{4}\left(\frac{1}{7} - \frac{0}{4}\right)^2 + \ldots$ it can be observed that the contribution to this distance of individual 1 is less than that of individual 2, in

spite of both having chosen option D1 and neither of them D4, which does not seem to be natural.

On the other hand, consider a category chosen by all the individuals in the survey. Imagine, for example, that in a public opinion survey all the individuals are men. In this case, variable gender is not relevant to the analysis since it does not enable any distinction to be drawn between opinions of men and of women. So the category 'man' (with a constant profile of $1/n$) should have no influence in the analysis of the survey. From the geometric point of view of the CA, this category would have to coincide with the centre of gravity of the cloud of categories and therefore there should be zero distance to this centre of gravity. This is indeed the case if CA is applied to the CDT: $d^2(j, G_J) = \sum_i \frac{1}{p_{i.}}\left(\frac{p_{ij}}{p_{.j}} - p_{i.}\right)^2 = \sum_i n\left(\frac{1}{n} - \frac{1}{n}\right)^2$, but in the analysis of the IDT the distance to the centre of gravity of a category chosen by all the individuals is not zero: $d^2(j, G_J) = \sum_i \frac{z}{z_{i.}}\left(\frac{1}{n} - \frac{z_{i.}}{z}\right)^2$ so that this distance has an influence on all the elements of the analysis: inertias, determination of factorial axes and so on.

It seems, therefore, that the direct application of standard CA is not appropriate for the study of an IDT.

4 Problems of Applying Classical CA to a Created CDT

Possible solutions to the above problem could get in the way of creating a CDT. It is therefore necessary to include new response categories. In the case of multiple response questions, it is necessary to create for each response category a complementary category that denies the previous one. Thus, if the question has J_q response categories, $2J_q$ categories will be finally analyzed. In the case of questions conditioned by a previous question, it is necessary to add for each question a dummy category indicating that respondents are not required to answer (NRA) that question. This means that as many dummy categories are added as there are conditioned questions. Moreover, if the conditioned question is a multiple response question, both types of artificial category have to be created. In this case, the J_q initial categories become $3J_q$ final categories. For example, in the case of Table 1, where the individuals who choose category C2 do not have to answer question D, it would be necessary to create 4 new categorical variables for question D, each of them with three categories of response: Di represents the original category, Di^C is the dummy category that denies the original one for those individuals who choose category C1 and it is coded as 0 for those individuals who choose category C2, since for them the dummy category Di-NRA has to be created.

This can be seen in Table 2 where, for the sake of simplicity, we only show questions C and D from Table 1. Question D is transformed into 4 categorical variables with 12 categories. Table 2 shows that the 4 'not required to answer' categories are identical and at the same time equal to category C2. Table 2 also shows that the dummy category Di^C only would be a null category and would not have influence

Table 2 Example of a CDT created from an IDT

	C1	C2	D1	$D1^C$	D1-NRA	D2	$D2^C$	D2-NRA	D3	$D3^C$	D3-NRA	D4	$D4^C$	D4-NRA
1	1	0	1	0	0	0	1	0	1	0	0	0	1	0
2	1	0	1	0	0	0	1	0	0	1	0	0	1	0
3	0	1	0	0	1	0	0	1	0	0	1	0	0	1
4	1	0	1	0	0	1	0	0	1	0	0	1	0	0
5	0	1	0	0	1	0	0	1	0	0	1	0	0	1
6	1	0	1	0	0	1	0	0	1	0	0	1	0	0
7	1	0	1	0	0	0	1	0	1	0	0	1	0	0
8	1	0	1	0	0	0	1	0	1	0	0	1	0	0
9	1	0	1	0	0	1	0	0	1	0	0	0	1	0
10	0	1	0	0	1	0	0	1	0	0	1	0	0	1

in the analysis if the answers of individuals to Di were identical to the answers to C1, as is the case with the category $D1^C$. A CDT created in this way facilitates the application of MCA, but it can also involve a number of problems. Since all the categories – both the original ones and the dummies – contribute to the creation of factorial axes, they give rise to planes covered by points, complicating the interpretation. Moreover, the interpretation has to be carried out cautiously since the additional categories may actually fit the negative of the original category but may also hide a desire not to answer and/or ignorance of the answer. The dummy categories can have identical or very similar response patterns and they can even create the first factorial axes, as is usually the case in conditioned questions (Zárraga and Goitisolo 2008). Therefore, the creation of a CDT and its analysis by means of MCA does not seem to be an adequate solution to the problem.

5 Proposed Methodology: CA of an IDT with a Modified Marginal

As seen above, the fact of having an IDT and therefore, a marginal on $\mathscr{I}$ depending on the different number of answers given by individuals can lead to inappropriate results. Consequently, a new methodology for analysing IDTs seems to be needed.

The methodology that we propose is based on CA with a modified marginal (Escofier 1987; Zárraga and Goitisolo 1999) and consists of replacing the marginal over $\mathscr{I}$ of the IDT which is not constant by an appropriate constant marginal in the whole analysis.

We propose that this marginal should be $1/n$. This choice is made for several reasons: (1) It is the natural centre of gravity of multiple choice questions and non conditioned question categories in the survey; (2) It gives equal weight or importance to all individuals; (3) The distances between individuals (Sect. 3) are due only to differences between their response patterns; (4) In the distances between two categories (Sect. 3) all individuals have the same importance, regardless of whether

they answer all the questions or not; (5) The categories chosen by all individuals, if this situation arises, do not influence the analysis and (6) It is not necessary to introduce dummy categories in the analysis.

The objective of CA of the IDT Z^*, with the imposed marginal, is to identify a low-dimensional subspace that fits the points as closely as possible using weighted least squares and then projecting orthogonally the points onto the subspace for visualization and interpretation. Let $\mathbf{D}_c$ be the diagonal matrix whose diagonal entries are column frequencies $p_{.j} = z_{.j}/z$ and let $\mathbf{D}_r$ be the diagonal matrix whose diagonal entries are the values of the imposed marginal. The row profiles of Z^*, with masses given by the diagonal of $\mathbf{D}_r$, set up the row-points cloud in an $\mathscr{J}$-dimensional Euclidean space, structured by the inner product and metric defined by the matrix $\mathbf{D}_c$. The column profiles of Z^*, with masses given by the diagonal of $\mathbf{D}_c$, set up the column-points cloud in an n-dimensional Euclidean space, structured by the inner product and metric defined by the matrix $\mathbf{D}_r$.

Let $g(j)$ be the projection of column j onto the subspace which comes closest to the set of column-points. The solution minimizes the following function: $\sum_j p_{.j} d^2(j, g(j))$ or equivalently maximizes the function: $\sum_j p_{.j} g^2(j)$ which is the variance (inertia in the context of CA) in the data table explained by the subspace.

For identifying the dimensions of the subspace, CA is based on decompositions of centered and normalized matrices, using either the eigenvalue-eigenvector decomposition of a square symmetric matrix or the singular-value decomposition (SVD) of a rectangular matrix. Following the SVD approach, CA of the incomplete disjunctive table Z^* can be carried out by calculating the SVD of the centered and normalized matrix $\mathbf{S}$: $\mathbf{S} = \mathbf{D}_r^{-1/2}\left[\frac{\mathbf{Z}^*}{z} - \mathbf{D}_r \mathbf{1}\mathbf{1}^T \mathbf{D}_c\right]\mathbf{D}_c^{-1/2} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ where $\boldsymbol{\Sigma}$ is the diagonal matrix with singular values in descending order: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_S > 0$ with S being the rank of matrix $\mathbf{S}$. If strict inequalities order the singular values, that is, there is no multiplicity of singular values, then the SVD is uniquely determined up to reflections in corresponding singular vectors. Otherwise, if two singular values are equal, their corresponding singular vectors are determined only up to rotations in their respective two-dimensional subspaces.

The squared singular values or eigenvalues $\sigma_s^2 = \lambda_s$ are called principal inertias in the context of CA and their sum $\sum_s \lambda_s$ is equal to the total inertia and equal to $trace(\mathbf{S}\mathbf{S}^T) = trace(\mathbf{S}^T\mathbf{S})$. In CA literature, the total inertia is the amount of the total variance in the data table. The columns of $\mathbf{U}$, called left singular vectors, and those of $\mathbf{V}$, the right singular vectors, are orthonormal, $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$.

Projections of rows, $\mathbf{f}_s$, and columns, $\mathbf{g}_s$, on the sth axis are calculated as principal coordinates: $\mathbf{f}_s = \sqrt{n}\,\mathbf{u}_s\sqrt{\lambda_s}$, $\mathbf{g}_s = \mathbf{D}_c^{-1/2}\mathbf{v}_s\sqrt{\lambda_s}$.

In classical CA, the transition relationships between projections of different points create a simultaneous representation that provides more detailed knowledge of the matter being studied. In CA with the imposed marginal, the transition relationships between $\mathbf{f}_s$ and $\mathbf{g}_s$ can be expressed as:

$$\mathbf{f}_s = n\left[\frac{\mathbf{Z}^*}{z} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{D}_c\right]\mathbf{g}_s\frac{1}{\sqrt{\lambda_s}} \qquad \mathbf{g}_s = \mathbf{D}_c^{-1}\left[\frac{\mathbf{Z}^*}{z} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{D}_c\right]^T\mathbf{f}_s\frac{1}{\sqrt{\lambda_s}}$$

6 Illustrative Example

For the sake of simplicity, we only present the analysis of one multiple response question without considering the rest of the questions in survey or the socio-demographic variables. In particular, we deal with a survey with multiple response questions via a classical CA of the CDT created from the IDT and via the methodology proposed. Both analyses were carried out with Splus. The classical CA of the IDT is not presented here, for the reasons set out in Sect. 3. The data analyzed come from the Survey on Households and the Environment conducted in 2008 by the Spanish Institute of Statistics. In this survey the question "Would you agree with the following measures for the protection of the environment?" has been chosen as an example. Information on nearly 27,000 people aged 16 and over is considered.

The question gives eight possible responses: Making separation of household waste obligatory; Restricting abusive consumption of water; Establishing a tax on fuel; Establishing restrictions on private transport; Establishing a tax on tourism; Installing renewable energy parks (windfarm, solar power) in your town in spite of the effect on the landscape; Paying more for alternative energy and Reducing traffic noise. The coding of the answers to this question in disjunctive form gives rise to an IDT. To analyze this table via classical MCA, each of the eight original response categories becomes a question, each one with two categories: the original and the dummy one (Sect. 4). These dummy categories create inertia in the cloud of points and take part in the determination of the factorial axes, as they would be considered as categories of active questions. This means that, as observed in Fig. 1a, the first two factorial axes are determined by these categories. More specifically, the first factorial axis only shows the opposition between the original response categories of the "created questions" and the dummy categories. This is logical because each question has only two response categories, but this phenomenon is not very interesting. The dummy categories also contribute to the second and subsequent axes. It is mainly in the fourth axis that we can find a relationship between the original response categories. This fourth axis (Fig. 1b) allows us to say that there is a difference between people in favour of paying more for alternative energy and those in favour of establishing a tax on tourism and in favour of establishing restrictions on private transport. However, this relationship is found in the first plane if we analyze the IDT with the proposed methodology (Fig. 1c). In this first plane we also see, in the first axis, a difference between people who agree to pay for the protection of the environment and those people not prepared to pay for it. Exploring further dimensions (Fig. 1d), we can find more information about the relationships between the data analyzed.

7 Discussion

The fact that there are multiple response questions and/or conditioned questions in surveys means that the table in which individual answers are coded in logical and disjunctive form is an incomplete table. The incorporation of dummy categories to

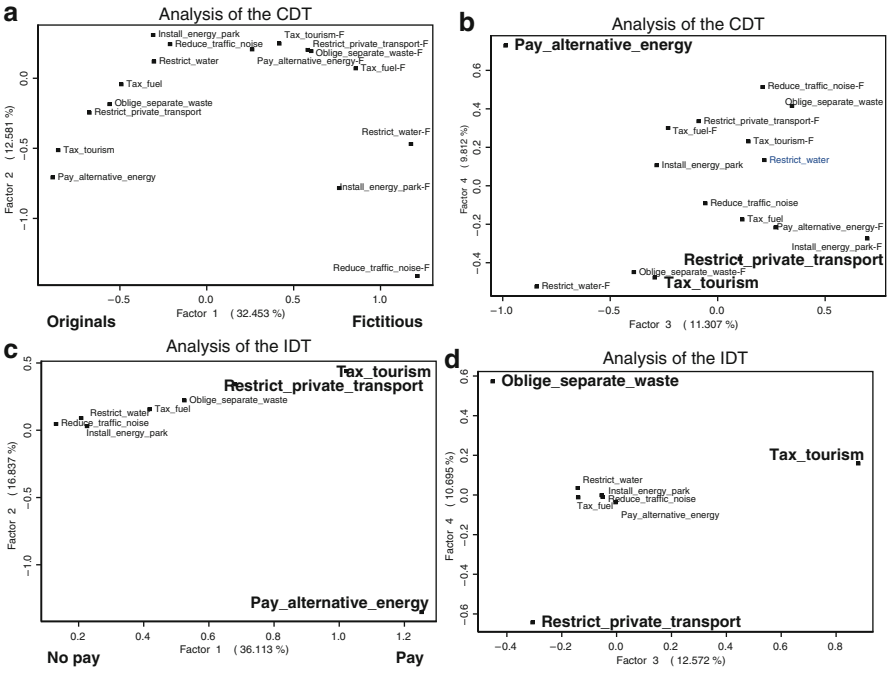


Fig. 1 Factorial planes

create a complete table can distort the analysis, since relationships between variables may appear which are due solely to the dummy categories. It therefore seems appropriate to work with incomplete disjunctive tables. However, the correspondence analysis of such tables reveals a defect in the application of the χ^2 distance to study the similarity between individuals. In IDT this distance increases not only with different answers from individuals but also with those categories chosen by both individuals if the number of questions answered does not coincide. The methodology proposed corrects this defect since the new χ^2 distance increases only with non common answers from respondents and allows us to look for the real relationships between questions or variables.

References

Escofier, B. (1987). Traitement des questionnaires avec non-réponse, analyse des correspondances avec marge modifiée et analyse multicanonique avec contrainte. *Publications de l' Institut de Statistique de l' Université de Paris*, *XXXII*(fasc 3), 33–70.

Escofier, B., & Pagès, J. (1998). *Analyses factorielles simples et multiples: Objectifs, méthodes et interprétation* (2nd ed.). Paris: Dunod.

Greenacre, M. J. (1984). *Theory and application of correspondence analysis*. London: Academic Press.

Lebart, L., Morineau, A., & Warwick, K. (1984). *Multivariate descriptive statistical analysis*. New York: John Wiley.

Zárraga, A., & Goitisolo, B. (1999). Independence between questions in the factor analysis of incomplete disjunctive tables with conditioned questions. *Questiió*, *23*(3), 465–488.

Zárraga, A., & Goitisolo, B. (2008). Análisis de encuestas con preguntas condicionadas. *Metodología de Encuestas*, *10*, 39–58.

Part XII
Multivariate Analysis

Control Sample, Association and Causality

Riccardo Borgoni, Donata Marasini, and Piero Quatto

Abstract We introduce the control sample in survey sampling as a tool for measuring the association between a possible cause X and a particular effect Y. The elements of the target population are divided in two groups according to whether X is present or not, the absence of X identifies the control group. A random sample is selected from each group with the aim of measuring the association between X and Y. We propose an unbiased estimator of the associational risk difference between groups and then generalize our approach to the problem of estimating the causal risk difference.

1 Introduction

In survey sampling theory from finite populations, the control sample can be introduced to measure the association between two dichotomous variables, denoted by X and Y. In this context, X may be a possible cause of the variable Y and the problem is that of measuring the strength of the supposed causal relation.

The finite population U can be divided into the two subgroups E of size N and C of size M by the variable X. In particular, E represents the group of primary interest ($X = 1$) and C the control group ($X = 0$).

To exemplify consider the case of quality evaluation of the university programs. In this context one relevant issue is to asses whether the type of secondary school certificate obtained by students affects the student drop-out later on in their university career. The population of students, U, is divided in the two groups, E and C, according to the type of secondary school certificate they obtained. For instance those who have got a technical diploma ($X = 1$) can be considered separately from the others ($X = 0$). In this case Y identifies those students who have dropped out from the university program after their first year ($Y = 1$). Hence, the problem is to measure the causal association between the student's drop out and having had a technical secondary school certificate.

In order to measure the fore-mentioned association, several measures have been introduced such as the relative risk,v the odds ratio and the risks difference. Each

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_58,

of these measures has some appealing properties and disadvantages (Lachin 2000; Agresti 2002; Jewell 2004; Hernan 2004). In particular, Hernan and Robins (2006) discussed how the relative risk can be used to estimate causal effect in the context of observational designs. Herein we focus on associational risk difference which is considered as a parameter d of the population U:

$$d = \bar{Y}_1 - \bar{Y}_0 = P(Y = 1|X = 1) - P(Y = 1|X = 0) \tag{1}$$

where $-1 \leq d \leq 1$.

Referring to the example, the two probabilities in (1) measure the risk of dropout in E and, respectively, in C. In particular, in the context of students' drop out or more generally in the field of quality evaluation, the relative risk and the odds ratio largely used in epidemiology are slightly unsatisfactory. For instance, knowing that the relative risk is halved when moving from one of the two groups to the other (e.g. students coming from two different type of schools) it is not particularly useful if the dropouts are respectively, say, 40 and 20 out of thousands of students. A similar argument applies to the odds ratio as well, at least for moderately rare events. The difference between proportions takes more in account the order of magnitude of the phenomenon considered, an information that seems hardly ignorable in public services evaluation.

In Sect. 2 we describe a suitable sampling design and we propose an unbiased estimator of the associational risk difference. Sections 3 and 4 provide an extension of our approach to the case of causal risk difference. Finally, a simulation study is presented in Sect. 5.

2 Sampling Design and the Associational Risk Difference Estimate

In order to estimate the parameter (1), we select a sample of size $t = n + m$ from $U = E \cup C$. Therefore, let s_1 be a sample of n units drawn from E (called the study sample), let s_0 be the control sample of m units drawn from C (called the control sample) and let

$$s = s_1 \cup s_0 \tag{2}$$

be the union of the two independent samples.

Suppose that sampling is without replacement and that the design gives a positive probability $p(s)$ only to the $\binom{N}{n}\binom{M}{m}$ samples s obtained by the following steps. The first step gives positive probabilities to $\binom{N}{n}$ study samples s_1 of size n selected from the subpopulation E. The second step assigns positive probabilities to $\binom{M}{m}$ control samples s_0 of size m selected from the subpopulation C. Let $p(s_1)$

be the probability of the sample s_1 and let $p(s_0)$ be the probability of s_0. Then the sampling design gives to the pooled sample (2) the probability $p(s) = p(s_1)p(s_0)$. Under simple random sampling without replacement (srswor) we have

$$p(s) = \frac{1}{\binom{N}{n}\binom{M}{m}}$$

and the proposed estimate of the parameter (1) is given by

$$\hat{d} = \sum_{i \in s_1} \frac{y_i}{n} - \sum_{j \in s_0} \frac{y_j}{m} = \hat{\bar{y}}_1 - \hat{\bar{y}}_0$$

where $\hat{\bar{y}}_1$ and $\hat{\bar{y}}_0$ are the usual estimates of the means $\bar{Y}_1$ and $\bar{Y}_0$, respectively. It is well known that the corresponding estimator $\hat{D}$ is unbiased under srswor and that

$$v(\hat{D}) = \frac{\hat{\bar{y}}_1(1-\hat{\bar{y}}_1)}{n-1}\frac{N-n}{N} + \frac{\hat{\bar{y}}_0(1-\hat{\bar{y}}_0)}{m-1}\frac{M-m}{M}$$

represents an unbiased estimate of the variance of $\hat{D}$

$$Var(\hat{D}) = \frac{\bar{Y}_1(1-\bar{Y}_1)}{n}\frac{N-n}{N-1} + \frac{\bar{Y}_0(1-\bar{Y}_0)}{m}\frac{M-m}{M-1}.$$

3 Causality

Following Hernan (2004) and Morgan and Winship (2007), let Y^1 be the potential outcome under the value $X = 1$ and let Y^0 be the potential outcome under the value $X = 0$. The variable X has a causal effect on the outcome variable Y if the causal risk difference

$$\delta = P(Y^1 = 1) - P(Y^0 = 1) \tag{3}$$

is not zero. In such a context, we give particular attention to the following two conditions: the consistency condition, given by

$$\forall x = 0, 1 \quad X = x \Rightarrow Y^x = Y, \tag{4}$$

meaning that the potential outcome Y^x is consistent with the actual outcome Y; the exchangeability condition, which means that the potential outcome is independent of X, i.e.

$$\forall x = 0, 1 \quad Y^x \coprod X. \tag{5}$$

Thus, under exchangeability (5)

$$P(Y^x = 1) = P(Y^x = 1 | X = x)$$

and under consistency (4)

$$P(Y^x = 1|X = x) = P(Y = 1|X = x),$$

so that

$$P(Y^x = 1) = P(Y = 1|X = x)$$

and then

$$\delta = P(Y^1 = 1) - P(Y^0 = 1) = P(Y = 1|X = 1) - P(Y = 1|X = 0) = d.$$

This means that, under consistency and exchangeability, the causal risk difference (3) equals the associational risk difference (1).

4 Estimation of the Causal Risk Difference

Following Hernan and Robins (2006), the exchangeability condition (5) is rarely met and the stratified design is a natural way to select a control sample similar to the study sample with the intention of weakening (5). For this purpose, we can introduce a suitable variable $S = 1, 2, \ldots, H$ and assume that exchangeability is true only within the strata defined by S:

$$\forall x = 0, 1 \quad \{Y^x \coprod X\}|S. \tag{6}$$

The validity of the causal inference strictly depends on the conditional exchangeability (6), no matter how many variables are included in the stratification.

Getting back to the problem of assessing whether the type of secondary school certificate is a potential cause of students' drop out sketched in Sect. 1, exchangeability implies that the probability of a potential withdrawing from university is the same irrespectively of the type of high school diploma actually received. This assumption is hardly believable in real situations. Conditional exchangeability, on the contrary, implies that this happens within the categories of a stratification variable which can be possibly derived by combining several student's characteristics such as age, gender or being a student worker, usually available from the University administration records. When considered within highly homogeneous profiles, exchangeability sounds to be a much more plausible situation.

By means of the stratification, we may consider the stratum-specific version of (1)

$$d_h = \bar{Y}_{1h} - \bar{Y}_{0h} = P(Y = 1|X = 1, S = h) - P(Y = 1|X = 0, S = h)$$

and then generalize (1) to the weighted mean

$$d = \sum_{h=1}^{H} w_h d_h = \sum_{h=1}^{H} w_h (\bar{Y}_{1h} - \bar{Y}_{0h}) \tag{7}$$

where

$$w_h = P(S = h) = \frac{N_h + M_h}{N + M}$$

represents the stratum weight and N_h is the number of those units having $X = 1$ while M_h is the number of those units having $X = 0$. Despite the fact that the literature does not seem to provide a clear interpretation of (7), we prove that this parameter is equal to the causal risk difference (3), if the assumptions of consistency and conditional exchangeability hold. Indeed, under condition (6)

$$P(Y^x = 1|S = h) = P(Y^x = 1|X = x, S = h)$$

and under consistency (4)

$$P(Y^x = 1|X = x, S = h) = P(Y = 1|X = x, S = h),$$

so that

$$P(Y^x = 1|S = h) = P(Y = 1|X = x, S = h)$$

and then

$$\begin{aligned}
\delta &= P(Y^1 = 1) - P(Y^0 = 1) \\
&= \sum_{h=1}^{H} P(S = h)[P(Y^1 = 1|S = h) - P(Y^0 = 1|S = h)] \\
&= \sum_{h=1}^{H} P(S = h)[P(Y = 1|X = 1, S = h) - P(Y = 1|X = 0, S = h)] = d.
\end{aligned}$$

We consider the stratified version of srswor, so that the plug-in estimator of (7) becomes

$$\widehat{D} = \sum_{h=1}^{H} w_h \widehat{D}_h = \sum_{h=1}^{H} w_h (\widehat{\bar{Y}}_{1h} - \widehat{\bar{Y}}_{0h}). \tag{8}$$

It can be proved that estimator (8) is unbiased and has variance

$$Var\left(\widehat{D}\right) = \sum_{h=1}^{H} w_h^2 \left[\frac{\sigma_{1h}^2}{N_h - 1} \left(\frac{N_h}{n_h} - 1 \right) + \frac{\sigma_{0h}^2}{M_h - 1} \left(\frac{M_h}{m_h} - 1 \right) \right] \tag{9}$$

with

$$\begin{cases} \sigma_{1h}^2 = \bar{Y}_{1h}(1 - \bar{Y}_{1h}) \\ \sigma_{0h}^2 = \bar{Y}_{0h}(1 - \bar{Y}_{0h}) \end{cases}. \tag{10}$$

Finally, we derive the optimal allocation by finding the minimum of (9) subject to the constraint

$$\sum_{h=1}^{H}(n_h + m_h) = t$$

where t represents the total sample size. For this purpose, the method of Lagrange multipliers provides the optimal sample sizes

$$\begin{cases} n_h = \frac{w_h \sigma_{1h}}{\sum_{k=1}^{H} w_k(\sigma_{1k}+\sigma_{0k})} t \\ m_h = \frac{w_h \sigma_{0h}}{\sum_{k=1}^{H} w_k(\sigma_{1k}+\sigma_{0k})} t \end{cases} \tag{11}$$

adopting for simplicity the commonly used approximations

$$N_h - 1 \approx N_h, M_h - 1 \approx M_h.$$

As usual, in order to realize the optimal allocation (11), it will be necessary to conduct a pilot survey for estimating the unknown parameters (10).

5 Simulation Study

In order to assess the loss of efficiency due to using the proportional allocation

$$\begin{cases} n_h = \frac{N_h}{N+M} t \\ m_h = \frac{M_h}{N+M} t \end{cases} \tag{12}$$

instead of (11), the results of a simulation study are reported herein. The simulation design is as follows. A finite population of $N = 1{,}500$ units and $M = 2{,}500$ units was generated. The population was stratified in two levels of a potential cause consisting of $N_1 = 450$ and $N_2 = 1{,}050$ and $M_1 = 1{,}500$ and $M_2 = 1{,}000$. For such a population, fixed values of d and d_h ($h = 1, 2$) were considered. Different scenarios for the associational risk difference were generated by changing the proportions $\bar{Y}_{1h}$ and $\bar{Y}_{0h}$ in each stratum. As reversing the sign leaves the results unaffected, (9) taking the same value, we do not report the case of both negative d_h in the table. The population was repeatedly sampled without replacement using both the proportional and optimal allocation for the strata. The plug-in estimator (8) was calculated for each sample. For the optimal allocation, we estimated the variances in (11) by calculating the plug-in counterpart of (10) on a simulated pilot survey of size equal to 1% of the population. The Monte Carlo (MC) mean, variance and design effect (i.e. the ratio between the Monte Carlo variance of the estimator in the case of the proportional and optimal design) were computed. The procedure was repeated for

Table 1 Results of a Monte Carlo study

d_1	d_2	d	Sampling fraction	Optimal allocation		Proportional allocation		Deff
				MC mean	MC variance	MC mean	MC variance	
0.60	0.20	0.40	0.05	0.3958	0.0034	0.3948	0.0039	1.15
			0.10	0.3936	0.0016	0.3960	0.0019	1.19
			0.20	0.3955	0.0007	0.3945	0.0008	1.14
			0.30	0.3950	0.0004	0.3951	0.0004	1.00
0.84	0.84	0.84	0.05	0.8399	0.0013	0.8402	0.0015	1.15
			0.10	0.8403	0.0007	0.8415	0.0008	1.14
			0.20	0.8402	0.0003	0.8401	0.0004	1.33
			0.30	0.8403	0.0002	0.8401	0.0002	1.00
−0.6	0.2	−0.19	0.05	−0.1871	0.0025	−0.1907	0.0038	1.52
			0.10	−0.1891	0.0012	−0.1901	0.0017	1.42
			0.20	−0.1902	0.0005	−0.1900	0.0008	1.60
			0.30	−0.1903	0.0003	−0.1915	0.0004	1.33

The design effect (Deff) is based upon 1,000 Monte Carlo replicates.

a range of d values and for a range of (increasing) sampling fractions. Results are reported in Table 1. We explored a large number of cases which looked extremely similar to those reported in Table 1. The two designs have similar behavior in terms of bias as expected because of the estimator (8) is unbiased. The optimal allocation seems to improve the performance of the estimator sensibly only when both the sample size and the risk difference are relatively small or when the stratum-specific risks differences have opposite sign.

References

Agresti, A. (2002). *Categorical data analysis*. Hoboken: Wiley.

Hernan, M. A. (2004). A definition of causal effect for epidemiological research. *Journal of Epidemiological Community Health, 58*, 265–271.

Hernan, M. A., & Robins, J. M. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiological Community Health, 60*, 578–586.

Jewell, N. P. (2004). *Statistics for epidemiology*. Boca Raton, FL: Chapman & Hall.

Lachin, J. M. (2000). *Biostatistical methods: The assessment of relative risks*. New York: Wiley.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: methods and principles for social research*. Cambridge: Cambridge University Press.

A Semantic Based Dirichlet Compound Multinomial Model

Paola Cerchiello and Elvio Concetto Bonafede

Abstract This contributions deals with the methodological study of a generative approach for the analysis of textual data. Instead of creating heuristic rules for the representation of documents and word counts, we employ a distribution able to model words along text considering different topics. In this regard, following Minka proposal, we implement a Dirichlet Compound Multinomial distribution that is a mixture of random variables over words and topics. Moving from such approach we propose an extension called sbDCM that takes into account the different latent topics that compound the document. The number of topics to be inserted can be known or unknown in advance, on the basis of the application context. Without losing in generality we present the case where the number and characteristics of topics are properly evaluated on the basis of available data.

1 Introduction

With the rapid growth of on-line information, text categorization has become one of the key techniques for handling and organizing data in textual format. Text categorization techniques are an essential part of text mining and are used to classify new documents and to find interesting information contained within several on-line web sites. Since building text classifiers by hand is difficult, time-consuming and often not efficient, it is worthy to learn classifiers from experimental data.

Many approaches have been proposed and tested within the text modelling context. Among them we can mention: the latent semantic analysis (LSA) (Deerwester et al. 1990), the probabilistic latent semantic analysis (pLSA) (Hofmann 1999), the latent Dirichlet allocation (LDA) (Blei and Lafferty 2006), the correlated topic model (CTM) (Blei et al. 2003) and finally the Independent Factor Topic Models (IFTM) (Putthividhya et al. 2009). All those models are considered generative approaches since they try to represent the word generation process by introducing suitable distributions, in particular the multinomial and the Dirichlet random variables. The more complicated version of those generative models introduces the concept of topics and the relative correlation among them.

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_59,

On the other hand is important to mention another interesting research path focusing on the burstiness phenomenon, that is the tendency of rare words, mostly, to appear in burst. The above mentioned generative models are not able to capture such peculiarity, that instead is very well modelled by the Dirichlet compound multinomial model (DCM). Such distribution was introduced by statisticians (Mosimann 1962) and has been widely employed by other sectors like bioinformatics (Sjolander et al. 1996) and language engineering (Mackay and Peto 1994). An important contribution in the context of text classification was brought by Minka (2003) and Madsen et al. (2005) that profitably used DCM as a bag-of-bags-of-words generative process. Similarly to LDA, we have a Dirichlet random variable that generates a multinomial random variable for each document from which words are drawn. By the way, DCM cannot be considered a topic model in a way, since each document derives specifically by one topic. That is the main reason why Doyle and Elkan (2009) proposed in 2009 a natural extension of the classical topic model LDA by plugging into it the DCM distribution obtaining the so called DCMLDA.

Following this path of research, we move from DCM approach and we propose an extension of the DCM, called 'semantic-based Dirichlet Compound Multinomial' (sbDCM), that permits to take latent topics into account.

2 Background: The Dirichlet Compound Multinomial

The DCM distribution (Madsen et al. 2005; Minka 2003) is a hierarchical model: on one hand, the Dirichlet random variable is devoted to model the Multinomial word parameters θ; on the other hand, the Multinomial variable models the word count vectors ($\bar{x}$) comprising the document. The distribution function of the DCM mixture model is:

$$p(\bar{x}|\alpha) = \int_{\theta} p(\bar{x}|\theta)\, p(\theta|\alpha) d\theta, \tag{1}$$

where $p(\bar{x}|\theta)$ is the Multinomial distribution:

$$p(\bar{x}|\theta) = \frac{n!}{\prod_{w=1}^{W} x_w} \prod_{w=1}^{W} \theta_w^{x_w} \tag{2}$$

in which $\bar{x}$ is the words count vector, x_w is the count for each word and θ_w the probability of emitting a word w; therefore a document is modelled as a single set of words ('bag-of-words'). The Dirichlet distribution $p(\theta|\alpha)$ is instead parameterized by α:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{w=1}^{W} \alpha_w)}{\prod_{w=1}^{W} \Gamma(\alpha_w)} \prod_{w=1}^{W} \theta_w^{\alpha_w - 1} \tag{3}$$

with $\alpha = \{\alpha_w\}$ the Dirichlet parameter vector for words; thereby the whole set of words ('bag-of-bags') is modelled. Thus a text (a document in a set) is modelled as

a 'bag-of-bags-of-words'. From another point of view, each Multinomial is linked to specific sub-topics and makes, for a given document, the emission of some words more likely than others. Instead the Dirichlet represents a general topic that compounds the set of documents and thus the DCM could be also described as 'bag-of-scaled-documents'.

The added value of the DCM approach consists in the ability to handle the 'burstiness' of a rare word without introducing heuristics (see Rennie et al. 2003). In fact, if a rare word appears once along a text, it is much more likely to appear again.

When we consider the entire set of documents (D), where each document is independent and identified by its count vector, ($D = \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_N\}$), the likelihood of the whole documents set (D) is:

$$\begin{aligned} p(D|\alpha) &= \prod_{d=1}^{N} p(\bar{x}_d|\alpha) \\ &= \prod_{d=1}^{N} \left(\frac{\Gamma(\sum_{w=1}^{W} \alpha_w)}{\Gamma(x_d + \sum_{w=1}^{W} \alpha_w)} \prod_{w=1}^{W} \frac{\Gamma(x_{dw} + \alpha_w)}{\Gamma(\alpha_w)} \right); \end{aligned} \tag{4}$$

where x_d is the sum of the counts of each word in the document d-th ($\sum_{w=1}^{W} x_{dw}$) and x_{dw} the count of word w-th for the document d-th.

3 Semantic-based DCM

As explained in Sect. 2, we have a coefficient (α_w) for each word compounding the vocabulary of the set of documents which is called 'corpus'. The DCM model can be seen as a 'bag-of-scaled-documents' where the Dirichlet takes into account a general topic and the Multinomial some specific sub-topics.

Our aim in this contribution is to build a framework that allows us to insert specifically the topics (known or latent) that compound the document, without losing the 'burstiness' phenomenon and the classification performance. Thus we introduce a method to link the α coefficients to the hypothetic topics, indicated with $\beta = \{\beta_j\}$, by means of a function $\alpha = F(\beta)$ which must be positive in β since the Dirichlet coefficients are positive. Note that usually $dim(\beta) << dim(\alpha)$ and, therefore, our proposed approach is parsimonious.

Substituting the new function into the integral in (1), the new model is:

$$p(\bar{x}|\beta) = \int_{\theta} p(\bar{x}|\theta)\, p(\theta|F(\beta))\, d\theta. \tag{5}$$

We have considered as function $F(\beta)$ a linear combination based on a matrix $\mathbf{D}$ and the vector $\bar{\beta}$. $\mathbf{D}$ contains information about the way of splitting among topics the observed count vectors of the words contained in a diagonal matrix $\mathbf{A}$ and $\bar{\beta}$ is a

vector of coefficient (weights) for the topics. More specifically we assume that:

$$\mathbf{A} = \begin{pmatrix} w_1 & & & \\ & \cdot & & \\ & & w_w & \\ & & & \cdot \\ & & & w_W \end{pmatrix}, \mathbf{D} = \begin{pmatrix} d_{11} & . & . & . & d_{1T} \\ . & . & & & . \\ . & & d_{wt} & & . \\ . & & & . & . \\ d_{W1} & . & . & . & d_{WT} \end{pmatrix}, \bar{\beta} = \begin{pmatrix} \beta_1 \\ . \\ \beta_t \\ . \\ \beta_T \end{pmatrix} D^* = \mathbf{A} \times \mathbf{D}$$

$$F(\beta) = D^* \times \bar{\beta} = \bar{\alpha} = \begin{pmatrix} \alpha_1 \\ . \\ \alpha_w \\ . \\ \alpha_W \end{pmatrix} \tag{6}$$

Note that:

- $\alpha_w = \sum_t^T d_{wt}^* \beta_t$, with T the number of topics;
- $d_{wt}^* = w_w \times d_{wt}$;
- d_{wt} is the coefficient for word w-th used to define the degree of belonging to topic t-th and by which a portion of the count of word w-th is assigned to that particular topic t-th.

By substituting this linear combination into formula (4), we obtain the same distribution but with the above mentioned linear combination for each α:

$$p(\bar{x}|\beta) = \frac{n!}{\prod_w^W x_w} \frac{\Gamma(\sum_w^W \sum_t^T d_{wt}^* \beta_t)}{\Gamma(\sum_w^W (x_w + \sum_t^T d_{wt}^* \beta_t))} \prod_w^W \frac{\Gamma(x_w + \sum_t^T d_{wt}^* \beta_t)}{\Gamma(\sum_t^T d_{wt}^* \beta_t)}; \tag{7}$$

This model is a modified version of the DCM, henceforth semantic-based DCM, and the new log-likelihood for the set of documents becomes:

$$log(p(D|\beta)) = \sum_d^N \left(log\Gamma(\sum_w^W \sum_t^T d_{wt}^* \beta_t) - log\Gamma(x_d + \sum_w^W \sum_t^T d_{wt}^* \beta_t) \right) + \\ + \sum_d^N \sum_w^W \left(log\Gamma(x_{dw} + \sum_t^T d_{wt}^* \beta_t) - log\Gamma(\sum_t^T d_{wt}^* \beta_t) \right) \tag{8}$$

In Fig. 1 we report the graphical representation of the new model where the α's are substituted by a function of the β's.

An important aspect of the proposed approach is represented by the number T of topics to be inserted into the semantic-based DCM that can be:

1. a priori known (i.e., fixed by field experts);
2. unknown, (i.e., to be estimated on the basis of the available data).

Without loss of generality we treat the case when T is unknown, leading to a random variable T (see Cerchiello and Bonafede 2009 for more details).

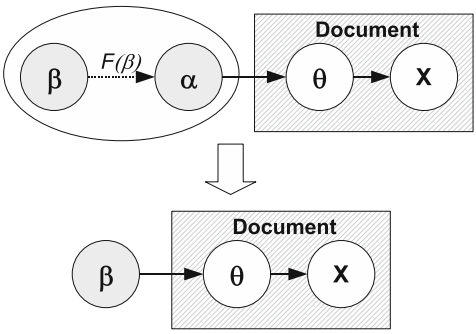


Fig. 1 Hierarchical model sbDCM representation

Since it is not always possible to know in advance the number of latent topics present in a corpora, it becomes very useful to build a statistical methodology for discovering efficiently and consistently a suitable number of topics.

In our context we tackle the problem as follows: in order to generate the coefficients contained within the matrix **D** we have used a segmentation procedure to group the words. The idea is to create groups of words sharing common characteristics that can be considered as latent topics. Such objective can be accomplished by applying a grouping analysis on the basis of the correlation matrix calculated on the complete set of words count. Later on, the analysis is completed by choosing the best number of groups and a distance matrix is used to set the membership percentage (d_{wt}) of each word to each latent topic. Since it is evidently reasonable, we allow words to belong to more than one topic; in fact it is likely that several topics can share common words. Thus the matrix **D** is inserted into formula (9) to find the coefficients β's and the new α's. Finally we use the α vectors into several classifiers based on discriminant rules.

4 Performance of the Semantic-Based Dirichlet Compound Multinomial

In this section we verify the goodness of the sbDCM models described in Sect. 3 , to understand if we can insert latent topics into DCM by maintaining the burstiness and the same classification performance.

Two kinds of matrixes have been used in the cluster procedure. One matrix contains the correlations among words (**C**) in the vocabulary and another one (**G**) is constructed by calculating the Kruskal-Wallis index (g) among words. The latter index g is defined as follows:

$$g = \frac{\frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(\bar{r}_i - \frac{N+1}{2}\right)^2}{1 - \frac{\sum_{i=1}^{p} (c_i^3 - c_i)}{N^3 - N}} \tag{9}$$

where n_i is the number of sample data, N the total observations number of the k samples, k the number of samples to be compared and $\bar{r}_i$ the mean rank of i-th group. The denominator of index g is a correction factor needed when tied data is present in the data set, where p is the number of recurring ranks and c_i is the times the i-th rank is repeated.

The index g depends on the differences among the averages of the groups ($\bar{r}_i$) and the general average. If the samples come from the same population or from populations with the same central tendency, the arithmetic averages of the ranks of each group ($\bar{r}_i = \sum_j r_{ij}/n_i$) should be similar to each other and to the general average $(N + 1)/2$ as well.

The training dataset contains 2,051 documents with a vocabulary of 4,096 words for both approaches. The evaluation dataset (again the same for both models) contains 1,686 documents which are distributed over 46 classes.

Once the α coefficients are obtained we employ three different classifiers described in Rennie et al. (2003) (normal (N), complement (C) and mixed (M)). All the classifiers select the document class with the highest posterior probability:

$$l(d) = argmax_c \left[log\ p(\theta_{\mathbf{c}}) + \sum_{w=1}^{W} f_w\ log\ \theta_{cw} \right] \tag{10}$$

where f_w is the frequency count of word w in a document, $p(\theta_{\mathbf{c}})$ is a prior distribution over the set of topics (that we consider uniformly distributed) and $log(\theta_{cw})$ is the weight for word w in class c.

The weight for each class is estimated as a function of the α coefficients:

$$\hat{\theta}_{cw} = \frac{N_{cw} + \alpha_w}{N_c + \sum_{w=1}^{N_c} \alpha_w} \tag{11}$$

where N_{cw} is the number of times word w appears in the documents belonging to class c, N_c the total number of words occurrences in class c.

In Tables 1 and 2 we report the results from the two tests obtained respectively on the basis of matrixes **C** and **G** and by varying the number of groups in the cluster analysis. In the tables we indicate with LL_{in} the log-likelihood before the parameters updating and with LL_{out} after the iteration procedure which is stopped when the predefined error ϵ (10^{-10}) is reached. The same with $AICc_{in}$ and $AICc_{out}$ that is the corrected Akaike Information Criterion (AICc) before and after the uploading. The indexes *Ind1*, *Ind2* and *Ind3* are defined as follows:

1. *(Ind1)* The proportion of true positive over the total number of test-documents:

$$\left(\sum_{d=1}^{D} \frac{TP_d}{D} \right) \times 100;$$

2. *(Ind2)* The proportion of classes that we are able to classify:

Table 1 Classification results by varying cluster numbers and using matrix (**C**)

Classifier	Measures	sbDCM_5	sbDCM_11	sbDCM_17	sbDCM_23	sbDCM_46	DCM
	LL_{in}	−282,226	−265,250	−252,197	−247,125	−242,991	−222,385
	LL_{out}	−205,412	−205,614	−205,601	−205,597	−205,602	−205,286
	$AICc_{in}$	564,462	530,522	504,228	494,296	486,074	454,264
	$AICc_{out}$	410,834	411,250	411,236	411,240	411,296	420,066
NORMAL	*Ind*1	68.13%	68.13%	67.95%	68.19%	68.13%	67.66%
//	*Ind*2	97.83%	97.83%	97.83%	97.83%	97.83%	97.83%
//	*Ind*3	62.32%	62.32%	62.15%	62.32%	62.32%	61.61%
COMP.	*Ind*1	68.19%	68.31%	68.25%	68.37%	68.25%	68.78%
//	*Ind*2	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
//	*Ind*3	66.20%	66.30%	66.05%	66.58%	66.01%	67.89%
MIXED	*Ind*1	68.07%	68.36%	68.43%	68.37%	68.31%	68.07%
//	*Ind*2	97.83%	97.83%	97.83%	97.83%	97.83%	97.83%
//	*Ind*3	63.87%	64.01%	64.05%	64.05%	64.01%	63.87%

Table 2 Classification results by varying cluster numbers and using matrix (**G**)

Classifier	Measures	sbDCM_5	sbDCM_11	sbDCM_17	sbDCM_23	sbDCM_46	DCM
	LL_{in}	−291,257	−283,294	−270,360	−266,453	−258,061	−222,385
	LL_{out}	−205,912	−204,647	−204,600	−204,604	−204,362	−205,286
	$AICc_{in}$	582,524	566,610	540,754	532,952	516,214	454,264
	$AICc_{out}$	411,834	409,316	409,234	409,254	408,816	420,066
NORMAL	*Ind*1	67.83%	67.71%	67.47%	67.42%	67.65%	67.66%
//	*Ind2*	97.83%	97.83%	97.83%	97.83%	97.83%	97.83%
//	*Ind3*	62.02%	61.73%	61.45%	61.43%	61.55%	61.61%
COMP.	*Ind1*	67.95%	68.66%	68.55%	68.72%	68.60%	68.78%
//	*Ind2*	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
//	*Ind3*	67.95%	68.09%	68.05%	68.29%	68.05%	67.89%
MIXED	*Ind1*	68.07%	68.13%	67.83%	67.71%	67.89%	68.07%
//	*Ind2*	97.83%	97.83%	97.83%	97.83%	97.83%	97.83%
//	*Ind3*	63.87%	63.97%	63.05%	62.86%	62.95%	63.87%

$$\left(\sum_{c=1}^{C} \frac{I_c}{C}\right) \times 100;$$

where I_c is an indicator that we set 1 if at least one document of the class is classified correctly, otherwise we set 0.

3. *(Ind3)* The proportion of true positive within each class over the number of test documents present in the class:

$$\left(\frac{1}{C}\sum_{c=1}^{C} \frac{TP_c}{M_c}\right) \times 100;$$

where M_c is the number of test-documents in each class, TP_c is the number of true positive in the class and C the number of topics (46).

As we can see in the two Tables 1 and 2, the percentages of correct classification (*Ind*1) are very close to the original ones with a parameter for each word (4,096 parameters). Of course they depend on the type of classifier employed during the classification step. Considering both sbDCM and DCM, the differences produced by varying the number of groups are small. Moreover the AICc is always better in the new approach instead of considering each word as a parameter as in the DCM model.

In particular for what concerns the approach based on the correlation matrix **C** (in Table 1) with 17 groups and on the Mixed classifier, it can predict correctly the 68.43% of documents. The log-likelihood and the AICc indexes along groups are quite similar, however the best value is obtained with 5 groups (respectively −205,412 and 410,834). Considering again the approach based on the correlation matrix **C**, we can conclude that, in terms of complexity expressed by the AIC index, the sbDCM approach, whatever applied classifier, is always better than the DCM.

When we use matrix **G** (see Table 2) the best classification performance is for the complement classifier based on 23 groups, with a percentage of 68.72%, a log-likelihood of −204,604, the AICc of 409,254. The best log-likelihood and AICc are for cluster with 46 groups (respectively −204,362 and 408,816). Even if the sbDCM distribution based on matrix **G** is not able to improve the classification performance of DCM, we can say that for sbDCM, Index one is always very close to the best one. In addition the new model is always better in terms of either AIC or log-likelihood indexes.

Moreover, if we perform an asymptotic chi-squared test (χ^2_{test}) considering the two cases (matrixes **G** and **C**) to decide whether the difference among log-likelihoods (LL), with respect to DCM, are significant (i.e., the difference is statistically meaningful if the $|LL_1 - LL_2|$ is greater than 6), we can see from Tables 1 and 2 the test with matrix **G** has the best performance.

Acknowledgements The paper is the result of the close collaboration between the authors. However Sects. 1, 2, 3 and 4 were written by Paola Cerchiello and the computational aspects were supervised by Elvio Concetto Bonafede. This work has been supported by MUSING 2006 contract number 027097, 2006–2010 and FIRB, 2006–2009.

References

Blei, D. M., & Lafferty, J. D. (2006). Correlated topic models. InWeiss, Y., Schölkopf, B., and Platt, J., (eds.), *Advances in Neural Information Processing Systems*, *18*, MIT Press, Cambridge, MA.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Cerchiello, P., & Bonafede, E. C. (2009). *Dirichlet compound multinomials for text modelling*. Technical Report available via http://www-3.unipv.it/dipstea/workingpapers.php.

Deerwester, S., Dumais, S., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent semantic analysis, *Journal of the American Society for Information Science*, *41*(6), 391–407.

Doyle, G., & Elkan, C. (2009). Accounting for burstiness in topic models. *Proceeding of International Conference on Machine Learning (ICML)*, *36*.

Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of Special Interest Group on Information Retrieval SIGIR*.

Mackay, D. J. C., & Peto, L. (1994). A hierarchical dirichlet language model. *Natural Language Engineering*, *1*(3), 1–19.

Madsen, R. E., Kauchak, D., & Elkan, C. (2005). Modeling word burstiness using the Dirichlet distribution. *Proceeding of the 22nd International Conference on Machine Learning*.

Minka, T. (2003). *Estimating a Dirichlet distribution*. Technical Report available via http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate B-distribution, and correlations among proportions. *Biometrika*, *49*(1 and 2), 65–82.

Putthividhya, D., Attias, H. T., & Nagarajan, S. S. (2009). Independent factor topic models. *Proceeding of International Conference on Machine Learning (ICML).*

Rennie, J. D. M., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifier. *Proceeding of the Twentieth International Conference on Machine Learning*.

Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I. S., & Haussler, D. (1996). Dirichlet mixtures: A method for improving detection of weak but significant protein sequence homology. *Computer Applications in the Biosciences*, *12*(4), 327–345.

Distance-Based Approach in Multivariate Association

Carles M. Cuadras

Abstract We show how to relate two data sets, where the observations are taken on the same individuals. We study some measures of multivariate association based only on distances between individuals. A permutation test is proposed to decide whether the association is significant. With these measures we can handle very general data.

1 Introduction

Many coefficients have been proposed in measuring the multivariate association between two random vectors or two data sets taken on the same individuals. Ecology is a clear example, where environmental data is related to species. In genomic data we may seek relations between genotype (e.g., DNA data) to phenotypes of interest. We can also find many examples in biometry and psychology, e.g., in relating physical characteristics to mental tests.

Often, the data sets are represented by two quantitative matrices $\mathbf{X}, \mathbf{Y}$ with the same number of rows, being the rows multivariate observations taken on the same individuals. Then some dependence measures based on canonical correlations can be used. However, if the data sets $\mathscr{D}_1, \mathscr{D}_2$ are non quantitative (binary, categorical, nominal), the information can alternatively be given by a similarity or a distance matrix. This distance-based approach, originated in Cuadras (1989), has been used as a tool in prediction and multivariate analysis, see Amat et al. (1998), Bartkowiak and Jakimiec (1994), Boj et al. (2007).

2 A First Multivariate Association Measure

Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ be a finite set with n individuals. Let $\delta_{ij} = \delta(\omega_i, \omega_j) = \delta(\omega_j, \omega_i) \geq \delta(\omega_i, \omega_i) = 0$ a distance or dissimilarity function defined on Ω. We suppose that the $n \times n$ distance matrix $\Delta_x = (\delta_{ij})$ is Euclidean. Then there exists

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_60,

a configuration $\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^p$, with $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$, $i = 1, \ldots, n$, such that $\delta_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)$. These coordinates of Ω constitute a $n \times p$ matrix $\mathbf{X} = (x_{ij})$ such that the distance between two rows i and j equals δ_{ij}

A well-known way of obtaining $\mathbf{X}$ from Δ_x finds $\mathbf{A} = -\frac{1}{2}\Delta_x^{(2)}$ and $\mathbf{G}_x = \mathbf{CAC}$, where $\Delta_x^{(2)} = (\delta_{ij}^2)$ and $\mathbf{C}$ is the centering matrix. The spectral decomposition $\mathbf{G}_x = \mathbf{U}\Lambda_x^2\mathbf{U}'$ provides $\mathbf{X} = \mathbf{U}\Lambda_x$. Matrices $\mathbf{X}$ and $\mathbf{U}$ contain the principal and standard coordinates, respectively, of the set Ω with respect to distance δ.

For a second data set on the same n individuals, we may consider another distance matrix Δ_y and find $\mathbf{G}_y = \mathbf{V}\Lambda_y^2\mathbf{V}'$. The principal coordinates are $\mathbf{Y} = \mathbf{V}\Lambda_y$ With these coordinates, the relation between the two data sets reduces to the relation between the centred matrices $\mathbf{X}(n \times p)$ and $\mathbf{Y}(n \times q)$.

Let us define the multivariate association measure between $\mathscr{D}_1, \mathscr{D}_2$ by

$$\eta(\mathscr{D}_1, \mathscr{D}_2) = \sqrt{\det(\mathbf{U}'\mathbf{V}\mathbf{V}'\mathbf{U})} = \sqrt{\det(\mathbf{V}'\mathbf{U}\mathbf{U}'\mathbf{V})}.$$

This measure satisfies $0 \leq \eta(\mathscr{D}_1, \mathscr{D}_2) = \eta(\mathscr{D}_2, \mathscr{D}_1) \leq 1$, and reduces to the multiple correlation coefficient when $\mathscr{D}_2$ is a data vector.

Let $\mathbf{S}_{xx} = \mathbf{X}'\mathbf{X}$, $\mathbf{S}_{xy} = \mathbf{X}'\mathbf{Y}$, $\mathbf{S}_{yy} = \mathbf{Y}'\mathbf{Y}$. As the (positive) canonical correlations $r_i, i = 1, \cdots, s = \min\{p, q\}$ between $\mathbf{X}$ and $\mathbf{Y}$ are the singular values of $\mathbf{S}_{xx}^{-1/2}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1/2} = \Lambda_x^{-1}\Delta_x\mathbf{U}'\mathbf{V}\Delta_y\Delta_y^{-1} = \mathbf{U}'\mathbf{V}$, this coefficient can be expressed as

$$\eta(\mathscr{D}_1, \mathscr{D}_2) = \prod_{i=1}^{s} r_i.$$

See Cuadras (2008) for further details. Since $\mathbf{U}'\mathbf{V}$ is a Gram matrix, measure η can be interpreted as the cosine of the angle between two subspaces expanded by $\mathbf{U}$ and $\mathbf{V}$, see Jiang (1996).

3 Multivariate Linear Regression

If we consider the columns of $\mathbf{X}$ and $\mathbf{Y}$ as predictor and response variables, respectively, a standard way to relate them is by multivariate linear regression

$$\mathbf{Y} = \mathbf{XB} + \Xi,$$

where $\mathbf{B}$ is a $p \times q$ matrix of parameters and Ξ is a $n \times q$ matrix of errors. The least-squares estimation of $\mathbf{B}$ is $\widehat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and the prediction matrix is $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\mathbf{B}} = \mathbf{PY}$ where $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the hat matrix.

Clearly, there is no relation if $\mathbf{B} = \mathbf{0}$. Assuming $\mathbf{X}, \mathbf{Y}$ centred, an appropriate statistic for deciding this null hypothesis is based on

$$F = [\mathrm{tr}(\mathbf{Y}'\mathbf{PPY})/(p+1)]/[\mathrm{tr}(\mathbf{Y}'(\mathbf{I}-\mathbf{P})\mathbf{Y})/(n-p-1)].$$

When $\mathbf{X}, \mathbf{Y}$ have been obtained by metric scaling from two distance matrices, then $\mathbf{X} = \mathbf{U}\Lambda_x$ and $\mathbf{P} = \mathbf{U}\mathbf{U}'$. As $\mathbf{P} = \mathbf{P}^2$ we have

$$\text{tr}(\mathbf{Y}'\mathbf{P}\mathbf{Y}) = \text{tr}(\mathbf{P}\mathbf{Y}\mathbf{Y}'\mathbf{P}) = \text{tr}(\mathbf{P}\mathbf{G}_y\mathbf{P}),$$

and similarly $\text{tr}[(\mathbf{Y}'(\mathbf{I}-\mathbf{P})\mathbf{Y}] = \text{tr}[(\mathbf{I}-\mathbf{P})\mathbf{G}_y(\mathbf{I}-\mathbf{P})]$. Therefore the ratio F can be formulated in terms of distances

$$\begin{aligned} F &= \frac{\text{tr}(\mathbf{P}\mathbf{G}_y\mathbf{P})/(p+1)}{\text{tr}[(\mathbf{I}-\mathbf{P})\mathbf{G}_y(\mathbf{I}-\mathbf{P})]/(n-p-1)} \\ &= \frac{\text{tr}(\mathbf{G}_x^-\mathbf{G}_x\mathbf{G}_y\mathbf{G}_x^-\mathbf{G}_x)/(p+1)}{\text{tr}[(\mathbf{I}-\mathbf{G}_x^-\mathbf{G}_x)\mathbf{G}_y(\mathbf{I}-\mathbf{G}_x^-\mathbf{G}_x)]/(n-p-1)}, \end{aligned}$$

where $\mathbf{G}_x^- = \mathbf{U}\Lambda_x^{-2}\mathbf{U}'$ is a g-inverse of $\mathbf{G}_x$.

We can invoke the F test when $q = 1$ and the only column of $\mathbf{Y}$ comes from a normal population. The F test is still justified when the rows of $\mathbf{Y}$ are multinormal with covariance matrix $\Sigma = \sigma^2\mathbf{I}$. For general data, testing $\mathbf{B} = \mathbf{0}$ can be performed by a permutation test.

To perform this test, we keep $\mathbf{Y}$ fix, then find the $n!$ permutations of the rows of $\mathbf{X}$ and obtain the permutation distribution of F. There will be evidence against $\mathbf{B} = \mathbf{0}$ if the observed F is in the extreme tail. If n is large, we may choose at random (with repetition) a subset of the $n!$ permutations.

Tests based on F when only $\mathbf{Y}$ comes from a distance, have been used by McArdle and Anderson (2001) in relating ecological data, and Wessel and Schork (2006) in large scale multilocus association studies. Here this test has been adapted to two distance matrices. However this F approach has four inconvenients. First, it depends on $\mathbf{G}_y = \mathbf{V}\Lambda_y^2\mathbf{V}'$, i.e., on the diagonal matrix Λ_y, whose entries are proportional to the variances of the columns of $\mathbf{Y}$. Second, Λ_y could have negative entries if the distance matrix is not Euclidean, as may occur in the presence of completely at random missing data. Third, if F is significant, we accept dependence but we do not know the degree of association between both data sets. Finally, F is non symmetric in $\mathbf{X}$ and $\mathbf{Y}$.

4 Measures of Multivariate Association

Another criterion used for testing $\mathbf{B} = \mathbf{0}$ in the multivariate linear regression model is the likelihood ratio criterion or Wilk's lambda, which is well known in multivariate analysis (Mardia et al. 1979). Wilk's lambda is

$$W = \det(\mathbf{E})/\det(\mathbf{E}+\mathbf{H}),$$

where $\mathbf{E} = \mathbf{Y}'\mathbf{Y} - \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}}$, $\mathbf{H} = \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}}$, with $\mathbf{Y} = \mathbf{V}\Lambda_y$ and $\widehat{\mathbf{Y}} = \mathbf{P}\mathbf{Y} = \mathbf{U}\mathbf{U}'\mathbf{V}\Lambda_y$. We then have $\mathbf{E} = \Lambda_y(\mathbf{I} - \mathbf{V}'\mathbf{U}\mathbf{U}'\mathbf{V})\Lambda_y$ and $\mathbf{E} + \mathbf{H} = \mathbf{Y}'\mathbf{Y} = \Lambda_y\mathbf{V}'\mathbf{V}\Lambda_y = \Lambda_y^2$. Therefore

$$W = \det(\mathbf{I} - \mathbf{V}'\mathbf{U}\mathbf{U}'\mathbf{V}).$$

Wilks' lambda does not depend on Λ_y and can be expressed in terms of canonical correlations

$$W = \prod_{i=1}^{s}(1 - r_i^2).$$

Clearly $A_W = 1 - W$ is also an association measure such that approaches 0 if $\mathbf{X}, \mathbf{Y}$ are independents, and approaches 1 if $\mathbf{X}, \mathbf{Y}$ are linearly dependent.

For testing $\mathbf{B} = \mathbf{0}$ we may also employ other criteria, such as Lawley-Hotelling and Pillai. If $\mathbf{H}\mathbf{v}_i = \lambda_i\mathbf{E}\mathbf{v}_i$ gives the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$, being $\lambda_i = r_i^2/(1 - r_i^2)$ here, the Lawley-Hotelling criterion U and Pillai's criterion V are

$$U = \mathrm{tr}[\mathbf{E}^{-1}\mathbf{H}] = \textstyle\sum_{i=1}^{s} r_i^2/(1 - r_i^2),$$
$$V = \mathrm{tr}[(\mathbf{E} + \mathbf{H})^{-1}\mathbf{H}] = \textstyle\sum_{i=1}^{s} r_i^2.$$

Then two multivariate measures of association can be based on $A_{LH} = (U/s)/(1 + U/s)$ and $A_P = V/s$. For the derivation of U, V and W, see Anderson (2003), Rao (1973).

The measure A_P also arises by applying the vectorial correlation between two random vectors $\mathbf{X}$, $\mathbf{Y}$, defined by (see Escoufier 1973):

$$RV = \mathrm{tr}(\mathbf{S}_{xy}\mathbf{S}_{yx})/\sqrt{\mathrm{tr}(\mathbf{S}_{xx}^2)\mathrm{tr}(\mathbf{S}_{yy}^2)},$$

where $\mathbf{S}_{xx} = \mathbf{X}'\mathbf{X}$, $\mathbf{S}_{xy} = \mathbf{X}'\mathbf{Y}$, etc. If we take standardized columns, then $\mathbf{S}_{xy} = \mathbf{U}'\mathbf{V}$, $\mathbf{S}_{xx} = \mathbf{U}'\mathbf{U} = \mathbf{I}_p$, $\mathbf{S}_{yy} = \mathbf{V}'\mathbf{V} = \mathbf{I}_q$, and this correlation reduces to $RV = (\sum_{i=1}^{s} r_i^2)/\sqrt{pq}$. Clearly $RV = A_P$ if $p = q$. However, if $p < q$ there are $q - p$ zero correlations, therefore $A_P = (\sum_{i=1}^{s} r_i^2)/s$ is better. In general $RV \neq A_P$.

Another measure of association, which generalizes the multiple correlation coefficient, is given by

$$R^2 = \det(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy})/\det(\mathbf{S}_{yy}).$$

Simplifying, we readily obtain $R^2 = \det(\mathbf{V}'\mathbf{U}\mathbf{U}'\mathbf{V}) = \prod_{i=1}^{s} r_i^2$, which we indicate by A_{HC} (Hotelling-Cramer, see Cramer and Nicewander 1979).

We can also relate $\mathbf{X}$ and $\mathbf{Y}$ via the Procrustes statistic (Cox and Cox 2001):

$$P^2 = 1 - [\mathrm{tr}(\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X})^{1/2}]^2/[\mathrm{tr}(\mathbf{X}'\mathbf{X})\mathrm{tr}(\mathbf{Y}'\mathbf{Y})].$$

In our distance-based context we obtain

$$P^2 = 1 - [\mathrm{tr}(\Lambda_x\mathbf{U}'\mathbf{V}\Lambda_y^2\mathbf{V}'\mathbf{U}\Lambda_x)^{1/2}]^2/[\mathrm{tr}(\Lambda_x^2)\mathrm{tr}(\Lambda_y^2)].$$

Standardizing the variables this equation reduces to

$$P^2 = 1 - [\text{tr}(\mathbf{U}'\mathbf{V}\mathbf{V}'\mathbf{U})^{1/2}]^2/(pq) = 1 - (\sum_{i=1}^{s} r_i)^2/(pq).$$

This measure suggests $A_{PR} = (\sum_{i=1}^{s} r_i)^2/s^2$.

Arenas and Cuadras (2004) proposed the association measure

$$A_{AC} = \text{tr}(\mathbf{G}_x^{1/2}\mathbf{G}_y^{1/2} + \mathbf{G}_y^{1/2}\mathbf{G}_x^{1/2})/\text{tr}(\mathbf{G}_x + \mathbf{G}_y),$$

which also lies between 0 and 1. Standardizing, from $\text{tr}(\mathbf{U}\mathbf{U}'\mathbf{V}\mathbf{V}') = \text{tr}(\mathbf{U}'\mathbf{V}\mathbf{V}'\mathbf{U})$, this equation reduces to

$$\begin{aligned} A_{AC} &= \text{tr}(\mathbf{U}\mathbf{U}'\mathbf{V}\mathbf{V}' + \mathbf{V}\mathbf{V}'\mathbf{U}\mathbf{U})/(p+q) \\ &= 2\text{tr}(\mathbf{U}'\mathbf{V}\mathbf{V}'\mathbf{U})/(p+q) \\ &= 2(\textstyle\sum_{i=1}^{s} r_i^2)/(p+q), \end{aligned}$$

which suggests $A_{AC} = (\sum_{i=1}^{s} r_i^2)/s = A_P$.

Finally Cramer and Nicewander (1979) give geometrical arguments to propose $A_{CN1} = A_{HC}^{1/s}$ and $A_{CN2} = 1 - W^{1/s}$. They also proposed the average $A_P = (\sum_{i=1}^{s} r_i^2)/s$.

Table 1 reports these measures in terms of $\mathbf{A} = \mathbf{V}'\mathbf{U}\mathbf{U}'\mathbf{V}$, $\mathbf{A}_c = \mathbf{I} - \mathbf{A}$, and $\lambda_i = r_i^2/(1 - r_i^2)$. Thus $W = \det(\mathbf{A}_c)$ and $A_P = \text{tr}(\mathbf{A})/s$.

Table 1 Some symmetric measures of multivariate association based on distances

Measure of association	Matrix expression in terms of $\mathbf{A} = \mathbf{V}'\mathbf{U}\mathbf{U}'\mathbf{V}$, $\mathbf{A_c} = \mathbf{I} - \mathbf{A}$.	Canonical correlation expression	Initially proposed by
r_1	First singular value of $\mathbf{U}'\mathbf{V}$	r_1	Hotelling
A_{HC}	$\det(\mathbf{A})$	$\prod_{i=1}^{s} r_i^2$	Hotelling-Cramer
A_W	$1 - \det(\mathbf{A}_c)$	$1 - \prod_{i=1}^{s} (1 - r_i^2)$	Wilks
A_{LH}	$[\text{tr}(\mathbf{A}_c^{-1}\mathbf{A})/s]/[1 + \text{tr}(\mathbf{A}_c^{-1}\mathbf{A})/s]$	$\frac{(\sum_{i=1}^{s} \lambda_i)/s}{1 + (\sum_{i=1}^{s} \lambda_i)/s}$	Lawley-Hotelling
A_P	$\text{tr}(\mathbf{A})/s$	$(\sum_{i=1}^{s} r_i^2)/s$	Pillai-Escoufier
A_{PR}	$[\text{tr}(\mathbf{A}^{1/2})]^2/s^2$	$(\sum_{i=1}^{s} r_i)^2/s^2$	Cox-Cox
A_{CN1}	$[\det(\mathbf{A})]^{1/s}$	$(\prod_{i=1}^{s} r_i^2)^{1/s}$	Cramer-Nicewander
A_{CN2}	$1 - [\det(\mathbf{A}_c)]^{1/s}$	$1 - [\prod_{i=1}^{s} (1 - r_i^2)]^{1/s}$	Cramer-Nicewander
η	$\sqrt{\det(\mathbf{A})}$	$\prod_{i=1}^{s} r_i$	Present author

5 Order Relationships

The above measures of association can be ordered as follows:

$$A_{HC} \leq \eta \leq A_{CN1},$$

and

$$A_{HC} \leq r_s^2 \leq A_{CN1} \leq A_{PR} \leq A_P \leq A_{CN2} \leq A_{LH} \leq r_1^2 \leq A_W,$$

where r_1^2 and r_s^2 are the largest and smallest squared canonical correlations. There is no order relationship between η and r_s^2.

Clearly $r_i^2 \leq r_i \leq r_i^{2/s}$, so

$$\prod_{i=1}^{s} r_i^2 \leq \prod_{i=1}^{s} r_i \leq (\prod_{i=1}^{s} r_i^2)^{1/s}$$

and

$$\prod_{i=1}^{s} r_i^2 \leq r_s^2 = (\prod_{i=1}^{s} r_s^2)^{1/s} \leq (\prod_{i=1}^{s} r_i^2)^{1/s} \leq (\sum_{i=1}^{s} r_i^2)/s,$$

as the geometric mean is less or equal to the arithmetic mean. We also have

$$[(\prod_{i=1}^{s} r_i)^{1/s}]^2 \leq [(\sum_{i=1}^{s} r_i)/s]^2 \leq (\sum_{i=1}^{s} r_i^2)/s.$$

Thus $A_{HC} \leq \eta \leq A_{CN1}$ and $A_{HC} \leq r_s^2 \leq A_{CN1} \leq A_{PR} \leq A_P$. For the other inequalities $A_P \leq A_{CN2} \leq A_{LH} \leq r_1^2 \leq A_W$, see (Cramer and Nicewander 1979).

6 Example and Discussion

As an illustration we use the data set from the low birth weight study, relating nine mixed variables, see Hosner and Lemeshow (2000). This data set contains information on 189 births and is available at http://www.umass.edu/statdata/statdata/. We relate two quantitative variables [Birth Weight, Weight of Mother] to seven quantitative, binary and categorical variables such as [Age], [Number of physician visits], [Smoking status], [History of hypertension], [Race], etc. Thus $p = 2, q = 7$, $n = 189$. We work with two distances: (a) Euclidean, (b) distance based on Gower's similarity coefficient.

For the Euclidean distance with $s = 2$ we obtain the canonical correlations $r_1 = 0.236, r_2 = 0.010$. For the Gower's distance with $s = 4$ we obtain the sequence of canonical correlations: $0.456, 0.272, 0.211, 0.008$.

The association measures are reported in Table 2. It is worth noting that: (1) all measures (except η) reduce to the squared multiple correlation coefficient when $\mathbf{Y}$

Table 2 Measures of multivariate association between birth weight, mother weight and seven quantitative, binary and categorical variables (age, number of physician visits, smoking status, race, etc.) in 189 births

Distance	A_{HC}	r_s^2	η	A_{CN1}	A_{PR}	A_P	A_{CN2}	A_{LH}	r_1^2	A_W
Euclidean	0.000	0.000	0.002	0.002	0.015	0.028	0.028	0.029	0.056	0.056
Gower	0.000	0.000	0.000	0.015	0.056	0.081	0.085	0.089	0.208	0.300

is univariate; (2) with the Euclidean distance we obtain the same results as classic canonical correlation analysis; (3) categorical variables (e.g., race) are non decomposed in dummy variables, as Gower's similarity coefficient is based on the number of matches for the multistate variables.

To decide whether A_W (Gower's distance) is significant, we perform a permutation test by keeping $\mathbf{V}$ fixed, finding the $n!$ permutations of the rows of $\mathbf{U}$ and obtaining the permutation distribution. The result $A_W = 0.3$ is in the extreme right tail of the distribution, so A_W is significant.

These measures have been obtained from distances, so we can relate two sets of mixed data, by using Gower's distance, as well as high dimensional data, by using the city-block distance, respectively. Whereas the classic canonical correlation fails when we have more variables than observations, this distance-based approach gives $r_1^2 = A_W = 1$. However the other coefficients may be less than 1, thus providing a non-trivial measure of association between two different data sets.

In general, these association measures can give values A_{HC} close to zero and A_W close to one. However, as it was reported by Rencher (1995), measures A_W, A_{CN2}, A_{LH} agree in general, but A_{HC}, RV, A_P may not indicate the same level of association, and further study is necessary to choose the most appropriate one for a given data set.

Acknowledgements Work supported in part by MEC (Spain) grant MTM2008-00642. Thanks are also due to an anonymous referee for useful comments.

References

Amat, L., Robert, D., Besalú, E., & Carbó-Dorca, R. (1998). Molecular quantum similarity measures tuned 3D QSAR: An antitumoral family validation study. *Journal of Chemical Information and Computer Sciences*, *38*, 624–631.

Anderson, T. W. (2003). *An introduction to multivariate analysis* (3rd ed.). New York: Wiley.

Arenas, C., & Cuadras, C. M. (2004). Comparing two methods for joint representation of multivariate data. *Communication in Statistics-Simultation and Computation*, *33*, 415–430.

Bartkowiak, A., & Jakimiec, M. (1994). Distance-based regression in prediction of solar flare activity. *Qüestiió*, *18*, 7–38.

Boj, E., Claramunt, M. M., & Fortiana, J. (2007). Selection of predictors in distance-based regression. *Communication in Statistics-Simultation and Computation*, *36*, 87–98.

Cox, T. V., & Cox, M. A. A. (2001). *Multidimensional scaling* (2nd ed.). Boca Raton: Chapman and Hall/CRC.

Cramer, E. M., & Nicewander, W. A. (1979). Some symmetric, invariant measures of multivariate association. *Psychometrika*, *44*, 43–54.

Cuadras, C. M. (1989). Distance analysis in discrimination and classification using both continuous and categorical variables. In Y. Dodge (Ed.), *Statistical data analysis and inference* (pp. 459–473). Amsterdam: Elsevier Science Publishers B. V.

Cuadras, C. M. (2008). Distance-based multisample tests for multivariate data. In B. C. Arnold, N. Balakrishnan, J. M. Sarabia, R. Mínguez (Eds.), *Advances in mathematical and statistical modeling* (pp. 61–71). Boston: Birkhauser.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, *29*, 751–760.

Hosner, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Jiang, S. (1996). Angles between Euclidean subspaces. *Geometriae Dedicata*, *63*, 113–121.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. London: Academic.

McArdle, B. H., & Anderson, M. J. (2001). Fitting multivariate models to community data: A comment on distance based redundancy analysis. *Ecology*, *82*, 290–297.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: Wiley.

Rencher, A. V. (1995). *Methods of multivariate analysis*. New York: Wiley.

Wessel, J., & Schork, N. J. (2006). Generalized genomic distance-based regression methodology for multilocus association analysis. *American Journal of Human Genetics*, *79*, 792–806.

New Weighed Similarity Indexes for Market Segmentation Using Categorical Variables

Isabella Morlini and Sergio Zani

Abstract In this paper we introduce new similarity indexes for binary and polytomous variables, employing the concept of "information content". In contrast to traditionally used similarity measures, we suggest to consider the frequency of the categories of each attribute in the sample. This feature is useful when dealing with rare categories, since it makes sense to differently evaluate the pairwise presence of a rare category from the pairwise presence of a widespread one. We also propose a weighted index for dependent categorical variables. The suitability of the proposed measures from a marketing research perspective is shown using two real data sets.

1 Introduction

Consider a general setup in which k categorical variables X_s $(s = 1, \ldots, k)$ with nominal scale are of interest and a categorical data set $\mathbf{X} = (\mathbf{x}_1', \ldots, \mathbf{x}_n')'$ is collected from n subjects $u_1, u_2, \ldots, u_n$. Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})'$ be the profile of the k attributes for the ith subject. The resemblance between two subjects u_i and u_j is typically measured by pairwise similarity indexes (see, e.g., Sneath and Sokal 1973 and, recently, Warrens 2008). Most of the available similarity indexes in the literature have been developed to deal with binary variables and few measures have been proposed specifically for polytomous attributes. For these variables, distance functions like the Euclidean or the Manhattan are sometimes used, especially for classification purposes. However, the principal difficulty in dealing with nominal categorical data is typically the lack of a metric space in which data points are positioned and the measured distances can be different when a different coding scheme is used for the variables (Zhang et al. 2006). In this paper we extend the original work of Zani (1982) and we first allow the classical similarity measures to deal with polytomous variables. Then we consider the problem of weighting variables in computing similarities between subjects. We propose a criterion for weighting the pairwise presence of a category, on the basis of the Shannon's "information content" of the relative frequency in the sample. Both in marketing research and in other fields, it appears relevant to attach a higher weight to the pairwise presence of a rare

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_61,

category in the sample rather than to the pairwise presence of a widespread one. A similar criterion is used in correspondence analysis, where the effect of increasing the values corresponding to low-frequencies categories relatively more than those corresponding to high-frequencies categories is accomplished with the χ^2 distance. Finally, we provide some numerical examples to illustrate the use of the indexes and we show the suitability of the proposed measures for market segmentation.

2 A Class of Similarity Indexes for Polytomous Variables

Consider k categorical variables X_s $(s = 1, \ldots, k)$ with $h_s \geq 2$ categories. An easy way to numerically code the attributes is through the so called "dummy variables". A binary variable is introduced for each category: the number of dummy variables is $h = \sum_{s=1}^{k} h_s$. With this coding scheme, we obtain a $(n \times h)$ data matrix of the form:

X_1	$\ldots$	X_s	$\ldots$	X_k
$X_{11} \ldots X_{1h_1}$	$\ldots$	$X_{s1} \ldots X_{sv} \ldots X_{sh_s}$	$\ldots$	$X_{k1} \ldots X_{kh_k}$
$\ldots$		$\vdots$ x_{isv} $\vdots$		$\ldots$
$\ldots$		n_{sv}		$\ldots$

where X_{sv} is the dummy variable for the vth category of the sth attribute ($s = 1, \ldots, k$ $v = 1, \ldots, h_s$). The observed value for the ith observation is $x_{isv} = 1$ if $x_{is} = v$ and $x_{isv} = 0$ if $x_{is} \neq v$. The frequency, in the sample, of the vth category of the sth attribute is $n_{sv} = \sum_{i=1}^{n} x_{isv}$ and the relative frequency is $f_{sv} = n_{sv}/n$. When the categorical variable is dichotomous, this coding scheme implies two dummy variables. It is obvious that $x_{is1} = 1 \iff x_{is2} = 0$ and $x_{is1} = 0 \iff x_{is2} = 1$ and the second dummy is superfluous. However, when dealing with mixed polytomous and dichotomous variables, the same coding is needed for both. To evaluate the similarity between subjects u_i and u_j, we introduce the following contingency table:

	1	0	tot
1	α	β	$\alpha + \beta$
0	γ	δ	$\gamma + \delta$
tot	$\alpha + \gamma$	$\beta + \delta$	h

We will call *positive matches* or *agreements* in u_i and u_j the α pairs $1 - 1$ and *disagreements* the $\beta + \gamma$ pairs $1 - 0$ and $0 - 1$. The δ pairs $0 - 0$ (*negative matches*) simply indicate that both u_i and u_j do not share the category corresponding to the dummy variable and are useless in evaluating the similarity between two subjects

since this number only depends on the number of the categories of the original categorical variables. The index:

$$S1_{ij} = \frac{\alpha}{\alpha + \beta + \gamma} \tag{1}$$

is bounded in [0,1] and has the following properties:

- $1 - S1_{ij} = (\beta + \gamma)/(\alpha + \beta + \gamma)$ is a distance. $\beta + \gamma$ is the Manhattan and the square Euclidean distance between u_i and u_j in the dummy variable coding.
- for binary variables, i.e., for $h_s = 2$, $s = 1, \ldots, k$, index $S1_{ij}$ becomes equivalent to the Rogers-Tanimoto index.

We may obtain a more general index by introducing a weight for the *disagreements* (Gower and Legendre 1986):

$${}_wS1_{ij} = \frac{\alpha}{\alpha + w(\beta + \gamma)} \tag{2}$$

with $w > 0$. When $w = 0.5$ and $h_s = 2$, $s = 1, \ldots, k$, expression (2) is equivalent to the Sokal-Michener index. Given two subjects u_i and u_j, the probability of an *agreement* in X_{sv}, in a Bernoulli trial, is f_{sv}^2. The weight given to an *agreement* in X_{sv} should be a decreasing function of f_{sv}^2. Assuming ($f_{sv}^2 > 0$), here we propose the weight $w_{sv} = log(1/f_{sv}^2)$ which is a measure of the information content of an *agreement* in X_{sv}. For independent variables, this measure is additive: if subjects u_i and u_j present two *positive matches*, in X_{sv} and X_{kl}, then the joint weight is $w_{sv,kl} = w_{sv} + w_{kl}$. The choice of w_{sv} is for interpretability purposes rather than for numerical ones. This weight is conceived in the light of the information theory. This criterion was first introduced by Burnaby (1970). We do not use the information entropy (see, e.g., MacKay 2002) $e_{sv} = f_{sv}^2 log(1/f_{sv}^2)$, because it is not a decreasing function of f_{sv}^2. By weighting the pairwise *positive matches* with w_{sv}, we obtain the index:

$$E_{ij} = \sum_{s=1}^{k} \sum_{v=1}^{h_s} \tau(i,j)_{sv} log\left(\frac{1}{f_{sv}^2}\right) \tag{3}$$

where $\tau(i,j)_{sv} = 1$ if $x_{isv} = 1$ and $x_{jsv} = 1$, $\tau(i,j)_{sv} = 0$ otherwise. $\sum_{s=1}^{k} \sum_{v=1}^{h_s} \tau(i,j)_{sv} = \alpha$. Expression (3) is equal to zero iff u_i and u_j do not share any *positive match*. However, it is not a similarity index since the condition $E_{ii} = E_{jj} = 1$ for $i, j = 1, \ldots, n$ is not satisfied. A general expression for a similarity index based on (3) is:

$$S2_{ij} = \frac{E_{ij}}{E_{ij} + F_{ij}} \tag{4}$$

with $F_{ij} \geq 0$ depending on the number of *disagreements* in u_i and u_j. We may define F_{ij} in different ways:

1. $F_{ij} = \sum_{s=1}^{k} \sum_{v=1}^{h_s} \sum_{t=1}^{h_s} \phi(i,j)_{svt} log\left(\frac{1}{f_{sv} f_{st}}\right)$ with $v \neq t$ and
   - $\phi(i,j)_{svt} = 1$ if $\begin{cases} x_{isv} = 1, x_{jsv} = 0 \;\&\; x_{ist} = 0, x_{jst} = 1 \\ x_{isv} = 0, x_{jsv} = 1 \;\&\; x_{ist} = 1, x_{jst} = 0 \end{cases}$
   - $\phi(i,j)_{svt} = 0$ otherwise.

   $\sum_{s=1}^{k} \sum_{v=1}^{h_s} \sum_{t=1}^{h_s} \phi(i,j)_{svt} = \beta + \gamma$
2. $F_{ij} = \sum_{s=1}^{k} \phi(i,j)_s log\left(\frac{1}{1-\sum_{v=1}^{h_s} f_{sv}^2}\right)$ and
3. $F_{ij} = \sum_{s=1}^{k} \phi(i,j)_s \sum_{v=1}^{h_s} f_{sv} log\left(\frac{1}{f_{sv}^2}\right)$ with
   - $\phi(i,j)_s = 1$ if $x_{isv} = 1, x_{jst} = 1$ and $x_{ist} = 0, x_{jsv} = 0$ $v \neq t$
   - $\phi(i,j)_s = 0$ otherwise.

   $\sum_{s=1}^{k} \phi(i,j)_s = 0.5(\beta + \gamma)$

In the first case, F_{ij} is equal to the information content of the specific pairwise *disagreements* in u_i and u_j and expression (4) becomes:

$$\frac{\sum_{s=1}^{k} \sum_{v=1}^{h_s} \tau(i,j)_{sv} log\left(\frac{1}{f_{sv}^2}\right)}{\sum_{s=1}^{k} \sum_{v=1}^{h_s} \tau(i,j)_{sv} log\left(\frac{1}{f_{sv}^2}\right) + \sum_{s=1}^{k} \sum_{v=1}^{h_s} \sum_{t=1}^{h_s} \phi(i,j)_{svt} log\left(\frac{1}{f_{sv} f_{st}}\right)} \tag{5}$$

Coefficient (5) is equal to (1) when $h_s = q$ and $f_{sv} = 1/q$, for $s = 1, \ldots, k$ and $v = 1, \ldots, q$. However, for two couples of subjects having the same pairwise *positive matches* but different pairwise *disagreements* it may assume a different value. Given the categorical nature of the data, the evaluation of the similarity should depend on the number of *disagreements* but not on the dummy variables in which they are present. In the second case, F_{ij} is the information content of any *dissimilarity*, without considering the specific dummy variable in which the *dissimilarity* is present. For attribute X_s, the probability of a *positive match* is $\sum_{v=1}^{h_s} f_{sv}^2$ and thus the probability of a *dissimilarity* is its complement $1 - \sum_{v=1}^{h_s} f_{sv}^2$. Considering this second expression, index (4) becomes:

$$\frac{\sum_{s=1}^{k} \sum_{v=1}^{h_s} \tau(i,j)_{sv} log\left(\frac{1}{f_{sv}^2}\right)}{\sum_{s=1}^{k} \sum_{v=1}^{h_s} \tau(i,j)_{sv} log\left(\frac{1}{f_{sv}^2}\right) + \sum_{s=1}^{k} \phi(i,j)_s log\left(\frac{1}{1-\sum_{v=1}^{h_s} f_{sv}^2}\right)} \tag{6}$$

Index (6) assumes the same numerical value for every pair of subjects having the same *positive matches*, regardless of the specific dummy variables in which the *disagreements* are present. In the trivial case in which $h_s = q$ and $f_{sv} = 1/q$ for $s = 1, \ldots, k$ and $v = 1, \ldots, q$, (6) becomes equal to (2) when $w = 0.5 log \frac{q}{q-1} / log(q^2)$. In the third case, F_{ij} is equal to the average of the information content of the pairwise *positive matches* in variables which have *disagreements* in u_i and u_j. Thus, F_{ij} may be perceived as the average loss in the information content due to the lack of

positive matches in u_i and u_j. For each attribute, the information content of a pairwise *positive match* in modality X_{sv} is $log(1/f_{sv}^2)$. The average of the information content is then $\sum_{v=1}^{h_s} f_{sv} log(1/f_{sv}^2)$. This quantity is also the average loss of the information content due to the lack of a *positive match* in X_s. With this expression:

$$S2_{ij} = \frac{\sum_{s=1}^{k} \sum_{v=1}^{h_s} \tau(i,j)_{sv} log\left(\frac{1}{f_{sv}^2}\right)}{\sum_{s=1}^{k} \sum_{v=1}^{h_s} \tau(i,j)_{sv} log\left(\frac{1}{f_{sv}^2}\right) + \sum_{s=1}^{k} \phi(i,j)_s \sum_{v=1}^{h_s} f_{sv} log\left(\frac{1}{f_{sv}^2}\right)} \tag{7}$$

Index (7) satisfies the following properties: (1) $S2_{ij} = 0$ iff u_i and u_j do not share any *positive match* and $S2_{ij} = 1$ iff u_i and u_j share a *positive match* in each attribute. (2) It is invariant to any permutation of the *disagreements*, provided that the *disagreements* are on the same attributes. (3) In the trivial case in which $h_s = q$ and $f_{sv} = 1/q$, for $s = 1, \ldots, k$, $v = 1, \ldots, q$, it becomes equal to (2) with $w = 0.5$. This last index, when $h_s = 2$, $s = 1, \ldots, k$, is equivalent to the Sokal-Michener measure. In order to take into account the possible association between variables, the information content E_{ij} of the pairwise *agreements* between two subjects should be defined in term of frequencies of specific 'sequences' of *agreements*. Let c_{ij} be the sequence of ones corresponding to pairwise *agreements* in u_i and u_j and $fr(c_{ij})$ be the relative frequency, in the sample, of observations holding the sequence c_{ij}. Consider, for example, three attributes having three, three and two categories, respectively. If the profile vectors of the dummy variables in u_i and u_j are $\mathbf{x}_i' = [001\ 100\ 01]$ and $\mathbf{x}_j' = [001\ 100\ 10]$, $fr(c_{ij})$ is the relative frequency, in the sample, of observations having the third category in the first attribute and the first category in the second attribute. Assume $fr(c_{ij})^2$ to be the probability, in a Bernoulli trial, of sampling two subjects with sequence c_{ij}. The information content of the sequence of *agreements* in u_i and u_j is: $L_{ij} = -2log\left(fr(c_{ij})\right)$ with the convention $L_{ij} = 0$ if u_i and u_j do not have any *positive match*. We normalize L_{ij} and we introduce the new similarity index:

$$S3_{ij} = L_{ij}/(L_{ij} + M(L_{ij})) \tag{8}$$

where $M(L_{ij})$ is the average information content of possible *agreements* in categories which have a *dissimilarity* in u_i and u_j. $M(L_{ij})$ may be thought of as the average loss in the information content due to the lack of pairwise *agreements*. Let us consider a number of *disagreements* equal to g, with $1 \leq g < k$. For $g = 0$, we introduce the convention $S3_{ij} = 0$. For $g = k$, $S3_{ij} = 1$. The count of observations having the same sequence of categories is $m_{ij} = n \times fr(c_{ij})$. In the sub sample of these m_{ij} observations, we determine the relative frequencies of each particular sequence of *agreements* for the remaining $(k - g)$ attributes having a *disagreements* in u_i and u_j. Since the number of categories in the sth attribute is h_s, the count of all possible sequences of *agreements* is $p = \prod_{s=1}^{k-g} h_s$ where the product is extended to the attributes having a *disagreement* in u_i and u_j. Let

– $c(ij)_t$ be the tth sequence of *agreements* among the $(k-g)$ attributes having a *disagreement* in u_i and u_j, $t = 1, \ldots, p$.
– $fr(c(ij)_t)$ be the relative frequency of the sequence $c(ij)_t$ in the sub sample of the m_{ij} subjects having the sequence c_{ij}.

The information content of $c(ij)_t$ is $-2log\,(fr(c(ij)_t))$. The average information content of the sequences $c(ij)_t$ is $M(L_{ij}) = \sum_{t=1}^{p} fr(c(ij)_t) - 2log\,(fr(c(ij)_t))$. With this expression, after algebraic simplification, index $S3_{ij}$ can be written as follows:

$$S3_{ij} = \frac{log\left(fr(c_{ij})\right)}{log\left(fr(c_{ij})\right) + \sum_{t=1}^{p} fr(c(ij)_t) log\,(fr(c(ij)_t))} \tag{9}$$

Using the conventions previously introduced, $S3_{ij} = 0$ iff u_i and u_j do not have any *agreement*, $S3_{ij} = 1$ iff u_i and u_j have a *positive match* in all k attributes.

3 Applications in Marketing Research and Discussion

In this section we try to gain insights into the characteristics of $S1$, $S2$ and $S3$ through applications in marketing research. All the analyses are performed in the Matlab environment (programs are available upon request). We also compare the proposed indexes with two popular similarity measures for polytomous variables: the Jaccard index (Sd) and the Hamming similarity index (Sh) (the function $I\{\cdot\} \in \{0, 1\}$ indicates the truth of its argument):

$$Sd_{ij} = \frac{I\{x_{is} = x_{js}, x_{is} \neq 0, x_{js} \neq 0\}}{I\{x_{is} \neq 0, x_{js} \neq 0\}} \qquad s = 1, \ldots, k \tag{10}$$

$$Sh_{ij} = \frac{I\{x_{is} = x_{js}\}}{k} \qquad s = 1, \ldots, k \tag{11}$$

While Sh is independent on the coding scheme, Sd depends on the code. For binary variables indicating the presence or the absence of a feature, we use the code `1` for the presence and `0` for the absence (so that the number of pairwise absences is not counted in Sd). For dichotomous variables in which the categories do not reflect the presence or the absence of a feature and for polytomous variables we use the code `1, 2, ..., s`. It is worth to highlight that Sd differs from $S1$ both in case of all polytomous variables and in case of mixed dichotomous and polytomous variables. We refer to Boriah et al. (2008) for a comparative study of the performances of a variety of similarity measures. The first data set consists of $k = 37$ observed features (technical specifications) of $n = 100$ satellite navigators. Seven variables have three categories and the other 30 variables are binary attributes (presence or absence). Some features, like a CD player, are very rare. Some other features, like a touch screen and a Gps system, are very common (see Table 1). Among the 666

Table 1 Relative frequencies of technical features in the satellite navigators

Attribute	fr	Attribute	fr	Attribute	fr
External slot	0.94	Slot expansion	0.92	Hard Disk	0.07
Mp3	0.58	Audio book	0.17	Picture viewer	0.61
In-built speaker	0.82	Camera	0.24	Automatic router	0.90
Autovelox	0.72	Bluetooth	0.44	Multimedia card	0.43
Touch screen	0.95	Gps	0.93	Internet connection	0.11
Cd player	0.02	Dvd	0.02	Tmb	0.60
Vocal control	0.05	Phone	0.18	Memory stick	0.11
Vocal warnings	0.91	In-built hard disk	0.27	Usb	0.76
Earpiece socket	0.91	Fm tuner	0.11	Tv tuner	0.16
IPod device	0.05	In-built antenna	0.97	High Memory	0.05

pairwise associations measured by the χ^2 statistics, there are 178 values leading to the rejection of the null hypothesis of independency between variables for $\alpha = 0.05$ and 110 for $\alpha = 0.01$. To analyze the behavior of the indexes, we consider the objects `Kenwood Dnx 7,200` (u_1) and `LG Lan 9,600 R` (u_2). They share the same category in 18 attributes (14 dichotomous and 4 polytomous). They both have a DVD and a CD player and the most rare category in two of the polytomous variables. The relative frequencies in the sample of the 18 categories shared by the objects are: 0.02, 0.24, 0.22, 0.03, 0.92 0.83, 0.61, 0.97, 0.91, 0.44, 0.95, 0.07, 0.02, 0.02, 0.6, 0.82 0.57, 0.99. The values of the similarity indexes are:

$$Sd_{12} = 0.38 \quad Sh_{12} = 0.48 \quad S1_{12} = 0.35 \quad S2_{12} = 0.69 \quad S3_{12} = 0.85$$

The Objects `Alpine Pmdb 100 P Blackbird` (u_3) and `Nokia E 61` (u_4) also have the same categories in 18 features (14 dichotomous and 4 polytomous). The relative frequencies, in the sample, of the 18 shared categories are: 0.92, 0.71, 0.75, 0.92, 0.91, 0.61, 0.91, 0.24, 0.9, 0.44, 0.93, 0.98, 0.98, 0.6, 0.95, 0.95, 0.82, 0.99. Since the number of *positive matches* and *negative matches* in the binary variables are also equal, the degree of similarity evaluated by Sd, Sh and $S1$ is identical:

$$Sd_{34} = 0.38 \quad Sh_{34} = 0.48 \quad S1_{34} = 0.35 \quad S2_{34} = 0.47 \quad S3_{34} = 0.70$$

but $S2_{34} < S2_{12}$ due to the pairwise presence of five very rare categories. For `Kenwood Dnx 7,200` (u_1) and `Qteck 9,090` (u_5) the values of the indexes are:

$$Sd_{15} = 0.55 \quad Sh_{15} = 0.59 \quad S1_{15} = 0.42 \quad S2_{15} = 0.85 \quad S3_{15} = 0.85$$

These satellite navigators share 22 categories (20 in binary variables and 2 in polytomous variables). Once again, the difference between $S1_{15}$ and $S2_{15}$ is due to the pairwise presence of the category `Ipod Interface`, which has a relative frequency in the sample equal to 0.05. While Sh and $S1$ are increasing functions of

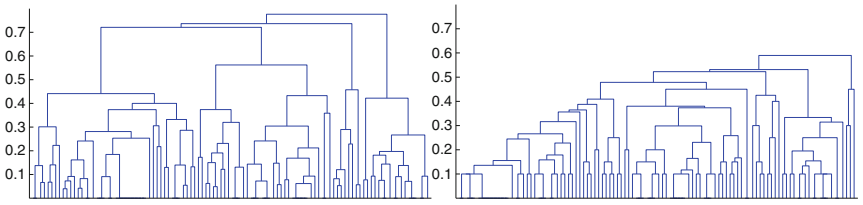


Fig. 1 Dendrograms obtained with the average linkage: in the *left* is used $S2$, in the *right* Sh

the number of shared categories, Sd and $S2$ may also decrease as the number of shared categories increase. Due to the weights, $S2$ is the most variable index. Sd may differ between couples of objects when the number of shared categories is equal but the number of pairwise absences differs. $S2$ may also differ when the number of shared categories and the number of pairwise absences are identical. Index $S3$ is not an increasing function of the number of *positive matches* since agreements in variables which are strongly associated are "penalized".

The second data set consists of $k = 10$ features of $n = 106$ sparkling wines. Five features are dichotomous variables and the others are polytomous. For this data set we obtain partitions with the most common hierarchical methods applied to the complements to one of the indexes. In marketing research, a specific criterion for assessing the performance of similarity measures is the 'segment addressability' suggested by Helsen and Green (1991), related to the degree to which a clustering solution can be explained by variables controlled by marketing managers and helping 'targeting' competitors. In Fig. 1, only two dendrograms are reported, for lack of space. In general, all classifications based on $S2$ and $S3$ readily distinguish four main groups while the other indexes show less ability to provide separation. Moreover, the classification based on $S2$ remains more stable, with respect to the different linkages, than those reached by the other measures. The four groups detected by $S2$ and $S3$ delineate specific segments of products and are easily interpretable also for the size (the smaller group comprehends eight wines and the biggest one 44 wines). These segments are homogeneous with respect to the alcohol content and the sugar level: the R^2 statistics for these variables is always higher in partitions reached with $S2$ and $S3$. Among the features used for classification, the 'taste' and the 'origin' have the rarest categories. In the 4-groups partition with $S3$, wines having the same modality in these two attributes are classified into the same group. With Sh, there is no evidence of a clustering structure in three or four groups: a very small cluster remains isolated until the last aggregation steps.

In conclusion, the major advantage of these indexes is that they are able to handle mixed dichotomous and polytomous variables and the weighted versions are able to give more importance to *agreements* in rare categories. Index $S3$ is designed to take into account the possible associations between variables and there do not appear to be other similarity measures that are directly focused on this goal.

References

Boriah, S., Chandola, V., & Kumar, V. (2008). Similarity measures for categorical data: a comparative evaluation. In *Proceedings of the 8th SIAM international conference on data mining*, Atlanta, pp. 243–254.

Burnaby, T. P. (1970). On a method for caracter weighting a similarity coefficient, employing the concept of information. *Mathematical Geology*, *2*, 25–38.

Gower, J. C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, *3*, 5–48.

Helsen, K., & Green, P. E. (1991). A computational study of replicated clustering with an application to marketing research. *Decision Science*, *22*, 1124–1141.

MacKay, D. J. C. (2002). *Information theory, inference and learning algorithms*, Cambridge, UK: Cambridge University Press.

Sneath, P. H., & Sokal, R. R. (1973). *Numerical taxonomy*. San Francisco, CA: Freeman.

Warrens, M. J. (2008). Bounds of resemblance measures for binary (presence/absence) variables. *Journal of Classification*, *25*, 195–208.

Zani, S. (1982). Sui criteri di ponderazione negli indici di similarità. In R. Leoni (a cura di) (Ed.), *Alcuni lavori di analisi statistica multivariata (*pp. 187–208). Firenze, Italia: SIS.

Zhang, P., Wang, X., & Song, P. X. (2006). Clustering categorical data based on distance vectors, *Journal of the American Statistical Association*, *101*, 355–367.

Causal Inference with Multivariate Outcomes: a Simulation Study

Paolo Frumento, Fabrizia Mealli and Barbara Pacini

Abstract Within the framework of the Rubin Causal Model, Principal Stratification is used to address post-treatment complications in randomized experiments, such as noncompliance, unintended missing outcomes, and truncation by death of the outcomes. We focus on a likelihood approach, exploiting the properties of multivariate finite mixture models in order to relax some of the usual identifying assumptions. These include monotonicity and exclusion restrictions hypotheses. A simulation study is conducted to show that the simultaneous modeling of more than one outcome may improve model identification and efficiency.

1 Introduction

Estimating causal effects of interventions is often the focus of empirical studies in medicine and social sciences. The only generally accepted approach for inferring causality requires that treatment receipt is randomized. Experiments, however, and social experiments in particular, often suffer from a number of complications, most notably noncompliance with assigned treatment, missing outcomes, and 'truncation by death' when the outcome is not always well-defined (e.g., quality of life when dead). The framework we adopt uses potential outcomes to define causal effects regardless of the mode of inference, often referred to as the Rubin Causal Model (RCM). Causal effects are defined by comparisons of potential outcomes on a common set of units (Rubin 1974, 2005). We apply Principal Stratification (PS; Frangakis and Rubin 2002), which was originally introduced to address post-treatment complications within the RCM.

We address the general issue of estimating the effect of a binary treatment on a multivariate outcome $\boldsymbol{Y} = \{Y_1, \ldots, Y_d\}$. We consider a very simple example, where only two outcomes (Y_1 and Y_2) are observed in a randomized experiment with noncompliance. We maintain the monotonicity of compliance and we assume that those assigned to the control group do not have access to the treatment. The population is then only composed by compliers (c) and never-takers (n). We relax the usual exclusion restriction on both outcomes. As showed in Angrist et al. (1996), without

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_62,

exclusion restrictions, additional assumptions are required in order to achieve point identification of treatment effects in the two groups. Imbens and Rubin (1997) and Hirano et al. (2000) show that relaxing the exclusion restriction(s) leads to 'weakly identified' models, with a proper but poorly informative, non-gaussian, even multimodal posterior distribution. Inference on causal effects is expected to be less accurate in absence of such strongly identifying assumption. We contribute to this existing literature by showing, in a wide simulation study, the potential advantages of the simultaneous modeling of two outcomes, in terms of model identification and efficiency. We adopt a likelihood approach and exploit the theory of Finite Mixture Models and the EM algorithm to obtain a point estimate of the causal parameters of interest.

2 Theoretical Framework

Suppose to design an experimental study, where a binary treatment is randomly assigned to a sample of N units; we denote with T_i $(i = 1, \ldots, N)$ the treatment assignment (1 if unit i is assigned to treatment, 0 if unit i is assigned to control), and with D_i the treatment receipt of unit i (1 if unit i receives treatment, 0 if unit i receives control). We thus assume that noncompliance is *all-or-none*. Noncompliance arises if $D_i \neq T_i$ for some i: if causal effects of treatment receipt on some (possibly multivariate) outcome $\boldsymbol{Y}_i$ are of interest, the estimation must take account of the nonrandom receipt of the treatment. Using the Principal Stratification (PS) approach (Frangakis and Rubin 2002), units can be classified into four principal strata, according to the compliance behavior. Using the potential outcomes notation, we denote with $D_i(t)$ and $\boldsymbol{Y}_i(t)$ the treatment receipt and the outcome of unit i, when assigned to t, $t = \{0, 1\}$. The following (latent) groups are defined:

- compliers $(c) = \{i : D_i(t) = t\}$,
- never-takers $(n) = \{i : D_i(t) = 0\}$,
- always-takers $(a) = \{i : D_i(t) = 1\}$,
- defiers $(d) = \{i : D_i(t) = 1 - t\}$.

A common monotonicity of compliance assumption rules out the presence of defiers, so that the proportion of defiers is equal to 0 ($P(d) = 0$). Very often, monotonicity of compliance holds by design, because those who are assigned to the control group do not have access to the treatment. This stronger monotonicity implies that $P(a) = P(d) = 0$. A more questionable assumption is that any effect of T_i on $\boldsymbol{Y}_i$ is through an effect of T_i on D_i (exclusion restriction, i.e., no assignment effect on a and n). With this set of assumptions, the average effect of receiving the treatment for compliers is point-identified and corresponds to the econometric instrumental variables (IV) estimand (Angrist et al. 1996).

Here we maintain the strong monotonicity assumption of compliance, so that the population is only composed by compliers (c) and never-takers (n). We instead relax the exclusion restriction: consequently causal parameters of interest are (a) the

average treatment effect for compliers, that is, the effect of treatment receipt, and (b) the average effect of treatment assignment for never-takers.

Under the stated assumptions, the actual (observed) likelihood function is:

$$L(\boldsymbol{\Phi}|\mathbf{T},\mathbf{D},\mathbf{Y}) = \prod_{i:T_i=1,D_i=1} \pi_i^c f_{i1c}(\mathbf{Y}_i) \prod_{i:T_i=1,D_i=0} (1-\pi_i^c) f_{i1n}(\mathbf{Y}_i) \prod_{i:T_i=0,D_i=0} \left(\pi_i^c f_{i0c}(\mathbf{Y}_i) + (1-\pi_i^c) f_{i0n}(\mathbf{Y}_i)\right), \quad (1)$$

where $\pi_i^c = P_i(c)$, while $f_{itc}(\mathbf{Y})$ and $f_{itn}(\mathbf{Y})$, $t = 0, 1$, are the generic joint density functions of the multivariate outcome $\mathbf{Y}$ for compliers and never-takers under treatment and under control. The first two factors in the likelihood represent the contribution of the compliers and never-takers assigned to the treatment group. The last factor represents the contribution to the likelihood function for those assigned to the control group. This includes both compliers and never-takers and the likelihood contribution is a mixture distribution. The role of the exclusion restriction is rather obvious: by imposing that $f_{i0n}(\mathbf{Y}) = f_{i1n}(\mathbf{Y})$ it would make the identification of the mixture components easier. In the absence on such an assumption, identification is more difficult (weak identification) and relies on the posed parametric assumptions.[1]

If we knew the compliance status for all sampled units, a different model could be fitted for each subpopulation; the ignorance of the true cluster membership for units assigned to control represents a source of uncertainty. The relevant issue is then how to accurately estimate this missing information. Improving the prediction of the unknown compliance status is generally expected to reduce the sampling variance and to speed the convergence of the EM algorithm. We now provide some intuitions on why a multivariate approach may improve the identification – and therefore the estimation – of a finite mixture model (Frumento 2009). We can think at each outcome variable as a “criterion" to assess the cluster membership of each sampled unit; because of this, the simultaneous modeling of all outcomes represents the way to make the “best" use of the sampling information. Conducting a separate analysis for each outcome variable has some advantages (low dimensionality, ease of implementation, opportunity of using standard routines in statistical packages), but also at least two main drawbacks, namely (a) it makes a partial use of the sampling information; and (b) it provides different estimates of the mixing proportions and of the unknown compliance status. In what follows, we provide some evidence on why a high-dimensional model is sometimes convenient in the analysis of finite mixture distributions.

[1] Covariates may also be included to predict the compliance behavior and the expected outcomes, according to some regression model; this generally improves the model identification.

3 Simulation Study

In order to have a realistic example in mind, suppose to design a social experiment to assess the efficacy of a language program for immigrants. The program is randomly offered to a sample of eligible secondary school students. Compliance is not perfect because only a subset of those assigned to treatment (i.e., offered to participate in the program) attend the classes; access to the program is denied to those assigned to control. The primary outcome of interest is language ability measured with the score achieved in a written test; in addition to the test result, also the time to complete the test is recorded and can be considered as a secondary outcome. Suppose these two outcomes, $\mathbf{Y} = (Y_1, Y_2)$, are distributed as a bivariate normal, conditional on compliance behavior and treatment assignment. If unit i is a complier,

$$Y_{i,1}, Y_{i,2} \sim N(\boldsymbol{\mu}_{c,i}, \boldsymbol{\Sigma}_c)$$

whereas if unit i is a never-taker,

$$Y_{i,1}, Y_{i,2} \sim N(\boldsymbol{\mu}_{n,i}, \boldsymbol{\Sigma}_n)$$

where

$$\boldsymbol{\mu}_{j,i} = \begin{pmatrix} \mu_{i,1:j} \\ \mu_{i,2:j} \end{pmatrix} = \begin{pmatrix} \alpha_{1:j} + \beta_{1:j} T_i \\ \alpha_{2:j} + \beta_{2:j} T_i \end{pmatrix} = \boldsymbol{\alpha}_j + \boldsymbol{\beta}_j T_i$$

$$\boldsymbol{\Sigma}_j = \begin{pmatrix} \sigma^2_{1:j} & \sigma_{12:j} \\ \sigma_{12:j} & \sigma^2_{2:j} \end{pmatrix}$$

for $j = \{c, n\}, i = 1, \ldots, N$. The model parameters are

$$\boldsymbol{\Phi} = \{\boldsymbol{\alpha}_c, \boldsymbol{\beta}_c, \boldsymbol{\Sigma}_c, \boldsymbol{\alpha}_n, \boldsymbol{\beta}_n, \boldsymbol{\Sigma}_n, \pi^c\},$$

where π^c is the proportion of compliers in the population. With this setting, $\beta_{1:c}$ and $\beta_{2:c}$ are the average treatment effects (ATE) on Y_1 and Y_2, respectively, in the subpopulation of compliers; they represent the causal effect of receiving the treatment; similarly, $\beta_{1:n}$ and $\beta_{2:n}$ represent the average effects of the treatment assignment on Y_1 and Y_2 in the subpopulation of never-takers. If the exclusion restriction holds on both outcomes, $\beta_{1:n} = \beta_{2:n} = 0$.

We selected a variety of values for the model parameters, the sample size, and the proportion of units assigned to the treatment; for each scenario, 2,000 simulated data sets have been used to compare three different models: (a) the univariate model for Y_1, (b) the univariate model for Y_2, and (c) the bivariate model for (Y_1, Y_2). We point out that estimating the couple of univariate models – rather than the bivariate one – has two main consequences: first, the covariance parameters $\sigma_{12:c}$ and $\sigma_{12:n}$ are not estimated; second, two different estimates of the mixing proportion π^c are obtained.

The standard EM algorithm with unconstrained covariance matrices has been used in the estimation; the two approaches have been compared according to three simple criteria: the MSE of the estimates of causal effects (relative MSE of the estimators of $\beta_{1:c}$ under the two approaches), the computation time (ratio between the mean computation time required in the estimation of the two univariate models, and of the bivariate one), and the prediction of the compliance behavior (ratio between the average proportion of correct group assignment, obtained by Bayes' classifier, using the univariate model for y_1, and the bivariate one).

Some results are reported in Tables 1–3. All statistics are referred to 2,000 Monte Carlo replications, with $\boldsymbol{\alpha}_c = (0, 0)$, $\boldsymbol{\alpha}_n = (0.2, 0.8)$, $\boldsymbol{\beta}_c = (0.3, 0.2)$, $\boldsymbol{\beta}_n = (0.05, 0.1)$, $\sigma_{1:c} = \sigma_{1:n} = 0.4$, $\sigma_{2:c} = \sigma_{2:n} = 0.8$, $\pi^c = 0.6$, $N = 300$ and $N_T = \sum T_i = 100$, and different combinations of the correlation coefficients ρ_c and ρ_n.

Simulations generally confirm our prior intuition; in most cases, the estimates obtained from the bivariate model have a smaller MSE and give a better prediction of the unknown compliance status. In addition, with respect to the univariate approach, noticeable gains in the computation time are observed. Stronger results are generally obtained when the number of units assigned to the treatment group (N_T) is small, so that the true compliance behavior (i.e., the cluster membership) is unknown for a great proportion of units.

The overlap of the two components of the mixture depends on the model parameters and affects the sampling distribution of the estimators; when the two distributions are strongly overlapping, the identification may be very weak. As we would expect, the bivariate approach is specially convenient in this case. In addition, the efficiency gain of the bivariate model when estimating treatment effects is generally larger for the outcome whose distributions are more overlapping.

Correlations ρ_c and ρ_n are additional parameters in the bivariate model and play a very important role. As the difference between ρ_c and ρ_n increases, the amount of

Table 1 Relative mean square error of the estimators of $\beta_{1:c}$ under the two approaches: $MSE(\hat{\beta}_{1:c,uni})/MSE(\hat{\beta}_{1:c,biv})$

$\rho_c \backslash \rho_n$	0	0.25	0.50	0.75
0	1.101	1.105	1.339	1.987
0.25	1.097	1.107	1.114	1.575
0.50	1.413	1.082	1.028	1.238
0.75	2.121	1.649	1.215	0.984

Table 2 Ratio between the mean computation time required in the estimation of the two univariate models, and of the bivariate one: $\bar{t}_{uni}/\bar{t}_{biv}$

$\rho_c \backslash \rho_n$	0	0.25	0.50	0.75
0	1.010	1.079	1.377	2.244
0.25	1.042	0.894	1.055	1.735
0.50	1.401	1.037	0.889	1.201
0.75	2.380	1.757	1.340	0.990

Table 3 Ratio between the average proportion of correct group assignment (among units assigned to the control condition) using the univariate model for y_1, and the bivariate model for y_1, y_2: $\bar{p}(g^*_{y_1} = g|T = 0)/\bar{p}(g^*_{y_1,y_2} = g|T = 0)$, where g^* denotes the predicted compliance behavior, obtained by Bayes' classifier

$\rho_c \backslash \rho_n$	0	0.25	0.50	0.75
0	0.882	0.874	0.858	0.817
0.25	0.894	0.904	0.891	0.845
0.50	0.869	0.901	0.916	0.882
0.75	0.808	0.838	0.862	0.902

information neglected by the univariate approach becomes more relevant. For this reason, the bivariate approach is specially convenient when compliers and never-takers are characterized by a very different value of ρ.

4 Concluding Remarks

Multivariate finite mixture models are applied in causal inference to address post-treatment complications. The simultaneous modeling of more than one outcome is generally expected to improve model identification. Even in case the causal effect of primary interest refers to a single outcome, modeling it jointly with other outcomes makes the estimation of the causal effect more efficient.

Using a simulation study, we evaluated the potential gains in efficiency, computation time, and predictive power obtained with a multivariate approach. Specifically, results show that the bivariate approach reduces the mean squared error of the estimates, speeding up the convergence of the EM algorithm and improving the prediction of the compliance behavior for units assigned to the control group. A similar reasoning can be applied in more complex settings, where a multivariate outcome is observed with arbitrary distributional assumptions.

We expect that some of these findings can be analytically proved and this is subject of our current research.

Our approach may be usefully implemented when the exclusion restrictions, or other identifying assumptions, are questioned; it can also be used to assess the robustness of the estimated treatment effects with respect to deviations from these assumptions, as a sort of sensitivity analysis of traditional IV estimates.

Appendix

We provide the steps of the EM algorithm (Dempster et al. 1977) used in estimating the model described in Sect. 2.

Under the stated assumptions, the actual (observed) likelihood function is reported in (1). The complete-data log-likelihood function can be written as:

$$
\begin{aligned}
l(\boldsymbol{\Phi}|\mathbf{T},\mathbf{D},\mathbf{Y}) = & \\
& \sum_{i:T_i=1,D_i=1} I(D_i(t)=t)[\ln \pi_i^c + \ln f_{i1c}(\mathbf{Y}_i)] \\
& + \sum_{i:T_i=1,D_i=0} I(D_i(t)=0)[\ln (1-\pi_i^c) + \ln f_{i1n}(\mathbf{Y}_i)] \\
& + \sum_{i:T_i=0,D_i=0} I(D_i(t)=t)[\ln \pi_i^c + \ln f_{i0c}(\mathbf{Y}_i)] \\
& + \sum_{i:T_i=0,D_i=0} I(D_i(t)=0)[\ln (1-\pi_i^c) + \ln f_{i0n}(\mathbf{Y}_i)],
\end{aligned}
$$

where $I(\cdot)$ is the general indicator function.

Once an initial value $\boldsymbol{\Phi}^{(0)}$ for the parameters has been chosen, the E-step of the EM algorithm computes the conditional expectation of the complete-data log-likelihood given the current vector of model parameters and the observed data. This corresponds to computing (using Bayes'rule) the conditional probabilities of the true compliance behavior indicator. The E-step is performed as follows:

$$
\begin{aligned}
&\text{for } i\colon T_i = 1, D_i = 1 \quad P(D_i(t)=t)^{(\tau)} = 1, \\
&\text{for } i\colon T_i = 1, D_i = 0 \quad P(D_i(t)=t)^{(\tau)} = 0, \\
&\text{for } i\colon T_i = 0, D_i = 0 \quad P(D_i(t)=t)^{(\tau)} = \frac{\pi_i^{c(\tau)} f_{i0c}^{(\tau)}(\mathbf{Y}_i)}{\pi_i^{c(\tau)} f_{i0c}^{(\tau)}(\mathbf{Y}_i) + (1-\pi_i^{c(\tau)}) f_{i0n}^{(\tau)}(\mathbf{Y}_i)}.
\end{aligned}
$$

The 'expected' log-likelihood function l_e is obtained by replacing the indicator functions with the computed probabilities of the true compliance behavior indicator in l: a new estimate of $\boldsymbol{\Phi}$, $\boldsymbol{\Phi}^{(\tau+1)}$, is then obtained using a standard optimization routine (M-step), as if $D_i(t)$ were known for all units.

As showed in Dempster et al. (1977), iterating this process monotonically increases the likelihood function, or at least leaves it unchanged; the algorithm runs until a stopping criterion has been satisfied.

References

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444–455.

Dempster, A. P., Laird, N., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data using the EM algorithm (with discussion). *Journal of the Royal Statistical Society. Series B*, *39*, 1–38.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, *58*, 21–29.

Frumento, P. (2009). *Finite mixture models. Some computational and theoretical developments with applications*. PhD Thesis, University of Florence.

Hirano, K., Imbens, G., Rubin, D. B., & Zhou, X. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, *1*, 69–88.

Imbens, G. W., & Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, *25*(1), 305–327.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non randomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. B. (2005). Causal inference using potential outcomes: design, modeling, decisions. *Journal of the American Statistical Association*, *100*, 322–331.

Using Multilevel Models to Analyse the *Context* of Electoral Data

Rosario D'Agata and Venera Tomaselli

Abstract Multilevel models are used to analyse contextual effects in hierarchical structures in order to explore the relationship among nested units. This study aims to observe the link among the territorial micro units nested in higher levels. We examine electoral data in two stages, defined in first level units inside nested structures. In these we used economic, demographic and social variables in order to characterize the context and explore its effects upon the electoral outline of territorial units.

1 Multilevel Modeling in the Analysis of Voting Behaviour

Hierarchical multilevel models have been used in scientific fields including the social sciences and have proved to be a methodological tool in many political behaviour studies, primarily of Anglo-Saxon origin (Steembergen and Jones 2002).

Since many subjective and contextual factors are involved in the direction of vote choice, it is important to highlight the role of context in affecting individual choices, thereby enhancing analysis and overcoming the restrictions set by classic methods like contextual analysis (Kreft and de Leeuw 1998).

The use of multilevel models have characterized studies meant to underline the importance of residential factors in political socialization processes (Cho et al. 2006), as well as surveys on voting intention (Barbosa and Goldstein 2000) and on ethnic group influence upon political participation.

One of the issues relating to voting behaviour is participation. In this case, hierarchical models have often been applied to choose contextual analysis units (Johnson et al. 2002) and to investigate context effects on abstentionism (Riba and Cuxart 2003). Other studies have underlined that contextual influences can not only derive from the territory where people reside and vote, but also from areas where they spend a great deal of their time, like for example, the labour place (Jones et al. 1992) and the household (Johnston et al. 2003).

The studies of political behaviour, with contextual elements as linking factors in explaining phenomena, take into consideration the territory as a relationship setting where individuals interact. Therefore, the territory is responsible for 'global effects' treated by Lazarsfeld and Menzel (1961).

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_63,

2 Materials and Methods

According to the above analysis framework, beyond the actual residence of electors, we can identify more aggregated territorial units, which link electoral tendencies to the contextual living area where opinions, attitudes and, therefore, patterns of behaviour are formed as a result of specific features of aggregated units.

By means of multilevel modelling, we analyse the electoral data of Italian Municipalities. As we did not choose to investigate individual behaviour drawn from survey data, Municipalities are used as level-1 units. Actually, we could have used polling station data, because this unit is the smallest, but in this case we could have employed only electoral data. Instead, we were interested in analysing the relationship between structural features and electoral outcomes. So, we chose, as level-1 unit, an aggregate-level unit (Goldstein 2003) – the Municipality, so as to enter socio-economical predictors in the model.

Multilevel modelling allows us to assess variation in a dependent variable at several levels simultaneously. They specify variables at any level, not just at the individual level (Kreft and de Leeuw 1998; Snijders and Bosker 1999). Here, we use the multilevel approach to analyse aggregated data in level-1 units.

So, this paper aims to compare two different hierarchical structures: Municipalities-provinces and Municipalities-Local Labour Systems. We propose the use of Local Labour Systems, called *Sistemi Locali del Lavoro* or *SLL*[1] in Italian, as level-2 aggregation units, to analyse voting behaviour by hierarchical linear models, with Municipalities as level-1 units.

The general model is:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \varepsilon_{ij}, \tag{1}$$

where:

$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01} z_j + u_{0j} \\ \beta_{1j} = \gamma_{10} + \gamma_{11} z_j + u_{1j}, \end{cases} \tag{2}$$

The Z_j, measured at level-2, is the contextual jth variable portion concerned with intercept and slope parameter measurement (Goldstein 2003; Snijders and Bosker 1999; Kreft and de Leeuw 1998). Thus, by introducing the Z_j variable of the Eq. 2, we use the model to estimate the $\beta_{0j}\beta_{1j}$ coefficients of (1).

[1] Introduced by Istat in 1981 and modified in number and structure during the 1991 and 2001 Census, the *SLLs* can be defined as "daily place of work and life activity". They are made up of Municipalities (comuni, in Italian) grouped on the basis of daily commuting to work as reported by Census data. Every area consists of several Municipalities. The majority of the resident population works within this area and employers recruit most of their man power from the Municipalities that make up this area (ISTAT 2001).

To analyse the voting behaviour of Municipalities (as level-1 units), we collected the 2006 national election data focusing on participation rate and the outcome of Centre-Left coalition. Since electoral results in a given area is strictly related to the same results in a previous election, we chose to employ the percentage of valid votes collected by the Centre-Left wing in the national election in 2001 as level-2 variable.

In order to fit a model able to examine the influence that territorial contexts would have on voting behaviour of hierarchically less aggregated units, structural dimensions were chosen on the basis of three factors of analysis: the socio-demographic dimension,[2] the socio-economic[3] dimension and the secularization dimension.[4]

3 Results

In the first step of analysis, we fitted a random intercept model (empty model) with the rate of consensus[5] obtained for the Centre-Left coalition in the 2006 national election (*Csnpol06*) as the dependent variable. We chose to model Centre-Left wing performance simply because it won that election. Considering the bipolar features of the Italian political system, furthermore, the performances of the two coalitions could be seen as mirror-like.

The electoral outcome was observed at the level-1 units: Municipalities. Then, we calculated the intra-class correlation coefficient (ICC) to assess the existence of meaningful differences between the level-2 units. ICC (ρ) is the proportion of total variance explained by level-2 unit (*SLLs*) variance. Formally:

$$\rho \in [0, 1] \tag{3}$$

where:

τ is the variance between level-2 units (*SLLs*)
σ^2 is the variance within level-1 units (Municipalities)

Comparing the two structures (Municipality-province and Municipality-*SLL*), we observed a greater homogeneity within the *SLL*s (Table 1). In fact, the ICC, calculated with *SLL*s level-2 units, shows a higher value (0.691) than the same coefficient calculated for the province level-2 units (0.566). In other words, Municipality voting behaviour would seem more influenced by nesting to a *SLL* than to a province.

[2] Rate of residents less than 5 years of age (*Meno_5*), the rate of families having over six members (*Fampiù6_2*), rate of foreign residents (*StranRes*), index of dependence (*Dependence*), rate of residents without qualifications (*NoTitStud*).

[3] Unemployment rate (*Disoccup_2*), youth unemployment rate (*DisocGiov*), the rate of families whose head is looking for work (*CapFamInCPO_2*), rate of workers in industry (*OccupInd_2*), rate of self-employed workers (*LavInPrp_2*), rate of housewives (*Casal_2*).

[4] Rate of unmarried couples (*NoConiug*).

[5] Computed as the ratio between the number of votes obtained by the Centre-Left coalition parties and the total number of valid votes.

Table 1 Empty model. Dependent variable: *Csnpol06*

	Municipalities-provinces				Municipalities-SLL			
	β_{0j}	E.S.	Z	p-Value	β_{0j}	E.S.	Z	p-Value
Fixed effects								
Intercept	49.221	0.945	52085.71	<0.0001	49.856	0.446	111784.75	<0.0001
Random effects								
u_{0j}	90.419	12.814	7.06	<0.0001	126.559	7.346	17.23	<0.0001
e_{0ij}	69.278	1.096	63.21	<0.0001	56.606	0.930	60866.67	<0.0001
ICC	0.566				0.691			

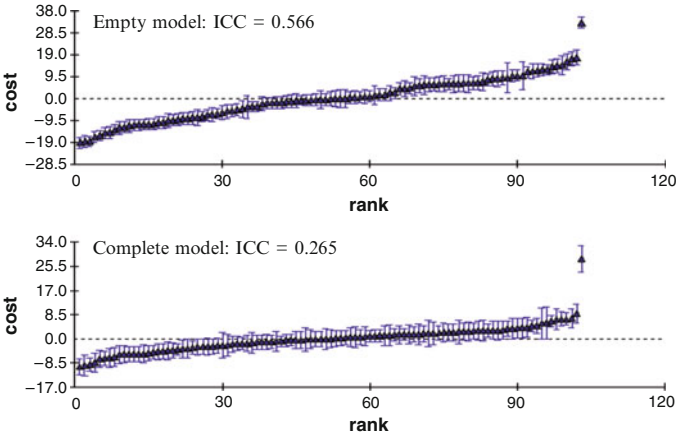


Fig. 1 Municipalities-provinces. Approximate 95% confidence interval for level-2 residuals plots: from empty model to complete model

Having fitted a random intercept model for all the independent variables and having observed the relationship between each one and the dependent variable, we estimated several models, assessing the effects of both level predictors.[6] Contextual effects were then assessed by introducing level-2 variables. Finally, after estimating the parameters of electoral predictors, we specified a model with all the predictors.

Before checking the value and direction of covariates in the model, let us observe the ICC decrease, comparing the two different structures (Fig. 1). In both hierarchical models with first and second level predictors, the ICC is lower than in the random intercept model. Namely, in the Municipality-province structure the ICC decreases from 0.566 to 0.265. In this case, controlling for the hierarchical structure by means of first and second level predictors, allowed us to point out the role of provinces in influencing Municipality electoral behaviour. Comparing second level residuals

[6] Primarily, we estimated a model with level-1 variables. Afterwards, assuming that the effects on the dependent would vary randomly across the level-2 units, we specified a random slope model to control interaction effects across the levels.

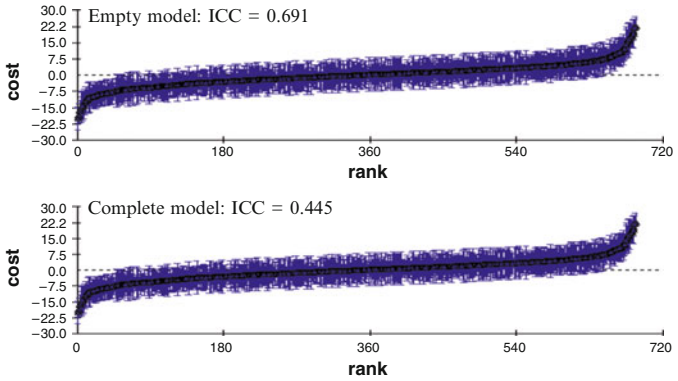


Fig. 2 Municipalities-*SLL*. Approximate 95% confidence interval for level-2 residuals plots: from empty model to complete model

plots, we note the reduction of variance due to nesting into Municipalities-provinces structure. In other words, the province seems to be a valid second level unit in predicting Municipality electoral outcome or, more probably, second level variables introduced in the model act in a wide territorial context. This entails a 'covering' of Municipalities structural peculiarities. With regards to Municipalities-*SLL* structure, we notice how the ICC decreases from 0.691 to 0.445 (Fig. 2). Considering the relative decrease, compared with Municipalities-provinces structure, the ICC is reduced by 64% whereas in the previous structure the ICC decreased, in relative value, by 46%.

Nevertheless, the explicative capability of second level variables in Municipalities-*SLL* structure seems to be lower than in the previous structure because ICC still maintains a high value after having estimated the complete model. The role of *SLL* nesting in influencing Municipalities electoral outcomes appears to be very strong and even controlling for the second level dimension, it was not possible to reduce the effects of nesting structure. Probably, other dimensions act on the dependent variable in this structure. Such a hierarchical structure, in fact, appears somewhat complicated and we need to introduce other predictors.

Furthermore, comparing the models of the two structures (Table 2) we note that in the Municipality-province model the level-2 variance is 24.003, while the empty model variance was 90.419. In the Municipality-*SLL* structure, the level-2 variance is 35.354 and was 126.559 in the random intercept model. So, in the Municipality-*SLL* model the predictors reduced the level-2 variance proportionally more than they did in the Municipality-province. In the first model, moreover, the predictors reduced the level-1 variance (from 50.606 to 44.041) more than in the Municipality-province structure (from 69.278 to 66.678).

Accordingly, *SLL* appears able to discriminate homogeneous nested territorial clusters more than the province does. Thus, Municipality-*SLL* structure maintains territorial peculiarities of Municipalities.

Table 2 Random intercept model: comparing two hierarchical structures. Dependent variable: *Csnpol06*

	Municipality-provinces				Municipality-SLL			
	β_{0j}	E.S.	Z	p-Value	β_{0j}	E.S	Z	p-Value
Level-1 fixed eff.(X_i)								
Intercept	11.231	5.407	2.077	<0.02	18.986	3.522	5.391	<0.0001
NoConiug	0.222	0.057	3.895	<0.0001	0.164	0.054	3.037	<0.002
Dependence	0.049	0.011	4.455	<0.0001	0.046	0.010	4.600	<0.0001
NoTitStud	−0.220	0.028	−7.857	<0.0001	−0.200	0.027	−7.407	<0.0001
StranRes	−0.283	0.060	−4.717	<0.0001	−0.277	0.060	−4.617	<0.0001
PartPol2006	0.169	0.025	7.840	<0.0001	0.164	0.024	6.833	<0.0001
AltPol2001	0.098	0.017	5.765	<0.0001	0.227	0.015	15.133	<0.0001
DisocGiov	0.020	0.010	2.000	<0.003	0.028	0.009	3.111	<0.0005
Meno_5	−0.809	0.109	−7.422	<0.0001	−0.736	0.105	−7.010	<0.0001
Level-2 fixed eff.(Z_j)								
Csnpol2001_2	0.680	0.061	11.148	<0.0001	0.675	0.028	24.107	<0.0001
OccupInd_2	−0.285	0.059	−4.831	<0.002	−0.263	0.032	−8.219	<0.0001
Casal_2	−0.238	0.098	−2.429	<0.006	−0.194	0.037	−5.243	<0.0001
Fampiù6_2	4.963	0.588	8.440	<0.0001	4.182	0.238	17.571	<0.0001
CapFamInCPO_2		1.595	0.179	<0.5	−1.797	0.566	−3.175	<0.0008
Disoccup_2	−0.271	0.339	−0.799	<0.3	−0.479	0.214	2.238	<0.02
LavInPrp_2	−0.175	0.127	−1.378	<0.1	−0.289	0.059	−4.390	<0.0001
Random effects								
u_{0j}	24.003	3.573	6.718	<0.0001	35.354	2.357	15.000	<0.0001
e_{0ij}	66.678	1.054	63.262	<0.0001	44.041	0.903	48.772	<0.0001
ICC	0.265				0.445			
Deviance	57335.4				56719.25			

The *SLL* hierarchical structure, therefore, reveals the impact of some level-2 effects which are not significant in the Municipality-province structure. The unemployment rate, the rate of families whose head is looking for work and the rate of self employed workers are measures related to Municipality features. Moreover, the variable with the greatest difference between the two hierarchical structures is the *Altpol2001* (the rate of votes obtained from non-coalition parties in 2001): 0.098 for the Municipality-province and 0.227 for Municipality-*SLL*. In other words, it is precisely this divergence that underlines the difference between the two structures. Otherwise the coefficients in the two structures appear similar in value and direction.

In detail, looking at Level-1 fixed effects we note that the rate of votes collected from Centre-Left coalition is positively related primarily to the rate of unmarried couples (*NoConiug*) and the ratio of electoral participation observed at the same election (*PartPol2006*) and negatively related to the rate of residents less than five years of age (*Meno_5*), the rate of residents without qualifications (*NoTitStud*) and the rate of foreign residents (*StranRes*).

Furthermore, concerning Level-2 fixed effects, we note that Centre-Left coalition success localises in those areas characterised by a high rate of families with

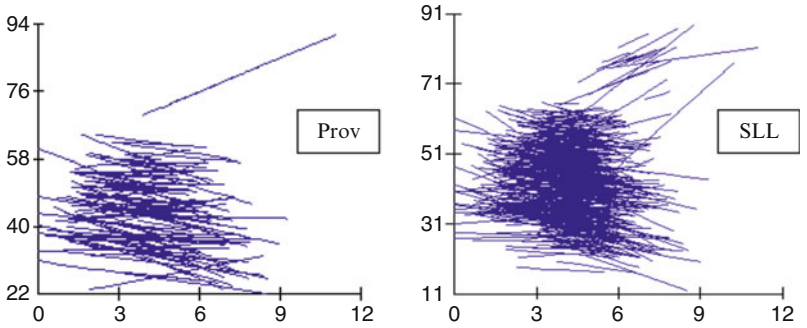


Fig. 3 Relationship between *Meno_5* and *Csnpol06* in two hierarchical structures (random slopes effects)

over six members (*Fampiù6_2*) and, of course, by the performance of Centre-Left parties observed in the previous election (*Csnpol2001_2*). Again, the consensus of the Centre-Left wing seems to be linked to the contexts featured by the low rate of workers in industry (*OccupInd_2*) and the low rate of housewives (*Casal_2*).

Municipality electoral data analysis should take into account the territorial features characterizing the context where the Municipalities are nested. Provinces, due to their extension and delimitation criteria, are less able to examine contextual features than *SLL*s, which could be more valid territorial units for the analysis of voting behaviour.

A central issue in electoral data analysis concerns the different capability of predictors in influencing the electoral outcome across all territorial units. In other words, many scholars are interested in investigating whether a given relationship between a covariate and outcome varies from unit to unit.

In our analysis, we answer this question. We propose estimating a random slope model in order to establish the different territorial effect of covariate on electoral outcome. Figure 3 describes an example of a randomly varying slope parameter for predictor *Meno_5*.

It shows that the strength of relationship between the covariate Meno_5 and the dependent Csnpol06 decreases in so far as the average of dependent variable increases. Every line represents level-2 units – provinces or SLLs – and their slopes indicate the different role of predictors across the territory.

4 Conclusions

The use of multilevel linear models has allowed us to identify the contextual factors involved in modelling electoral outcomes of Municipalities. These were observed in two hierarchical structures, defined according to different criteria. A random intercept model emphasized, in particular, the effects of the hierarchical structure on Italian Municipality voting behaviour. By estimating these models, it emerges that

*SLL*s have a greater classifying capability than provinces do and reveal features and analytic dimensions which the Municipality-province hierarchical structure does not reveal.

To conclude, from the analysis it appears that *SLL*s more than provinces, allow us to study the link between the context and voting behaviour. The latter is not always related to the two major coalitions and is often linked to the territorial/local dimension.

References

Barbosa, M. F., & Goldstein, H. (2000). Discrete response multilevel models for repeated measures: an application to voting intention data. *Quality and Quantity, 34*, 323–330.

Cho, W. K. T., Gimpel, J. C., & Dyck, J. J. (2006). Residential concentration, political socialization and voter turnout. *The Journal of Politics, 68*(1), 156–167.

Goldstein, H. (2003). *Multilevel statistical models*. London: Hodder Arnold.

ISTAT. (2001). *14° Censimento della popolazione e delle abitazioni*. Roma: ISTAT.

Johnson, M., Philips Shively, W. & Stein, R. (2002): Contextual data and the study of elections and voting behavior: Connecting individuals to environments. *Electoral Studies, 21*, 219–233.

Johnston, R., Jones, K., Sarker, R., Burgess, S., Propper, C., & Bolster, A. (2003). A missing level in the analysis of British voting behaviour: the household as context as shown by analyses of 1992–1997 longitudinal survey. PSA EPOP Conference, Cardiff, UK.

Jones, K., Johnston, R., & Pattie, C. J. (1992). People, places and regions: exploring the use of multi-level modelling in the analysis of electoral data. *British Journal of Political Science, 22*, 343–380.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel models*. London: Sage.

Lazarsfeld, P. F., & Menzel, H. (1961). On the relation between individual and collective properties. In A. Etzioni (Ed.), *Complex organization*. New York: Holt.

Riba, C., & Cuxart, A. (2003). Associationism and electoral participation: a multilevel study of 2000 Spanish general election. Comunicaciòn presentada en el VI congreso de la Asociaciòn española de ciencia polìtica y de la administraciòn. In *Capital social, Asociacionismo y participaciòn polìtica en España*. Barcelona, 18–20 de Septiembre.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: an introduction to basic and advanced multilevel modelling*. Thousand Oaks, CA: Sage.

Steembergen, M. R., & Jones, B. S. (2002). Modelling multilevel data structures. *American Journal of Political Science, 46*(1), 218–237.

A Geometric Approach to Subset Selection and Sparse Sufficient Dimension Reduction

Luca Scrucca

Abstract Sufficient dimension reduction methods allow to estimate lower dimensional subspaces while retaining most of the information about the regression of a response variable on a set of predictors. However, it may happen that only a subset of the predictors is needed. We propose a geometric approach to subset selection by imposing sparsity constraints on some coefficients. The proposed method can be applied to most existing dimension reduction methods, such as sliced inverse regression and sliced average variance estimation, and may help to improve the estimation accuracy and facilitate interpretation. Simulation studies are presented to show the effectiveness of the proposed method applied to two popular dimension reduction methods, namely SIR and SAVE, and a comparison is made with LASSO and stepwise OLS regression.

1 Introduction

Consider the regression of a response Y on a vector X of p predictors. Dimension reduction methods aim at finding a subspace $\mathscr{S}$ of minimal dimension such that

$$Y \perp\!\!\!\perp X | (\beta_1{}^\top X, \ldots, \beta_d{}^\top X) \tag{1}$$

where $\perp\!\!\!\perp$ denotes statistical independence, and $\boldsymbol{B} = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^{p \times d}$, with $d \leq p$, is the matrix spanning the basis of the subspace $\mathscr{S}$. Such subspace exists and is unique under mild conditions (Cook 1998a). We refer to it as the central dimension reduction subspace (CDRS) and denote it by $\mathscr{S}_{y|X}$.

Several methods have been proposed for estimating the basis $\boldsymbol{B}$ of the CDRS. These methods include sliced inverse regression (SIR, Li 1991), sliced average variance estimation (SAVE, Cook and Weisberg 1991), principal Hessian directions (PHD, Li 1992, Cook 1998b), and inverse regression estimation (IRE, Cook and Ni 2005).

All methods produce linear combinations of all the original predictors. Thus, unless a subset of predictors has exact null coefficients on all the directions, all the features are included. This often makes it difficult to interpret the extracted

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_64,

components. Furthermore, some predictors may have small coefficients, so their contribution is negligible.

As a simple illustration, consider the response model $Y = \exp(-0.75\beta^\top X + 1) + \epsilon$, where X is a vector of length $p = 6$, and both the predictors and the error ϵ are generated from independent standard normal variables. The central subspace is spanned by $\beta = (1, -1, 0, 0, 0, 0)^\top/\sqrt{2}$. For a simulated datasets of 200 observations SIR yields the estimate $\widehat{\beta} = (0.71, -0.694, 0.064, 0.082, -0.043, 0.044)^\top$. Although the coefficients for the last four predictors are nearly zero, all the variables are included in the estimate, and this obscures the fact that the true direction only involves the first two predictors.

A rough subset selection procedure could be based on the approximate formula for standard deviations of SIR estimates proposed by Chen and Li (1998). A rigorous conditional independence test procedure has been developed by Cook (2004) to assess the contribution of individual predictors to SIR components. Ni et al. (2005) and Li and Nachtsheim (2006) proposed a lasso (Tibshirani 1996) estimator to obtain sparse SIR estimates. Li (2007) generalized this approach to produce sparse estimates in dimension reduction methods based on a generalized eigenvalue formulation.

In this paper we propose an approach to sparse basis estimation based on a measure of distance between the estimated basis and a candidate sparse basis, followed by a penalized criterion for selecting relevant predictors. Search algorithms are also discussed.

2 Method

Given n independent realizations $\{(X_i, y_i),\ i = 1, \ldots, n\}$ of (X, y), let $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ be the $n \times p$ data matrix, and $y = (y_1, \ldots, y_n)^\top$ the $n \times 1$ response vector. Dimension reduction methods provide the estimate $\widehat{\boldsymbol{B}}$ of the $(p \times d)$ basis of the CDRS. In this paper we consider the structural dimension $d = \dim(\mathscr{S}_{y|X})$ as known.

Let $\widetilde{\boldsymbol{B}} = (\widetilde{\beta}_1, \ldots, \widetilde{\beta}_d)$ be the $(p \times d)$ sparse basis obtained from $\widehat{\boldsymbol{B}}$ by fixing at zero some coefficients for a given subset of predictors. Such sparsity constraint implies that if a predictor has zero coefficients on all the directions then it can be dropped from the active set of predictors. A similar approach was discussed by Chipman and Gu (2005) in the context of principal components analysis.

A Measure of Closeness of Subspaces

Following Li (1991), a measure of the distance between the estimated basis $\widehat{\boldsymbol{B}}$ of the CDRS and its sparse version $\widetilde{\boldsymbol{B}}$ is given by the squared trace correlation. This is based on an affine-invariant discrepancy measure which evaluates the distance of a direction $\widetilde{\beta}_j$ $(j = 1, \ldots, d)$ and the subspace $\mathscr{S}(\widehat{\boldsymbol{B}})$:

$$R^2(\widetilde{\beta}_j) = \max_{\widehat{\beta} \in \mathscr{S}(\widehat{\boldsymbol{B}})} \frac{(\widetilde{\beta}_j{}^\top \widehat{\boldsymbol{\Sigma}} \widehat{\beta})^2}{(\widetilde{\beta}_j{}^\top \widehat{\boldsymbol{\Sigma}} \widetilde{\beta}_j)(\widehat{\beta}^\top \widehat{\boldsymbol{\Sigma}} \widehat{\beta})} = \max_{\widehat{\beta} \in \mathscr{S}(\widehat{\boldsymbol{B}})} \operatorname{cor}(\boldsymbol{X}\widetilde{\beta}_j, \boldsymbol{X}\widehat{\beta})^2, \tag{2}$$

where $\widehat{\boldsymbol{\Sigma}}$ is the sample covariance matrix of the predictors. The *squared trace correlation* is defined as the average of the squared canonical correlation coefficients in (2) between the estimated directions $\boldsymbol{X}\widehat{\beta}_1, \ldots, \boldsymbol{X}\widehat{\beta}_d$ and the sparse directions $\boldsymbol{X}\widetilde{\beta}_1, \ldots, \boldsymbol{X}\widetilde{\beta}_d$:

$$R^2(\widetilde{\boldsymbol{B}}) = \frac{\sum_{j=1}^d R^2(\widetilde{\beta}_j)}{d}. \tag{3}$$

If $d = 1$ the squared trace correlation is equivalent to the square of the usual correlation coefficient. It can be shown that $R^2(\widetilde{\boldsymbol{B}}) = \cos(\theta)^2 \in [0, 1]$, where θ is the minimal angle between the subspaces $\widehat{\boldsymbol{B}}$ and $\widetilde{\boldsymbol{B}}$ with respect to $\widehat{\boldsymbol{\Sigma}}$ inner product. Thus, $R^2(\widetilde{\boldsymbol{B}})$ provides a measure of the "closeness" of the two subspaces: if $R^2(\widetilde{\boldsymbol{B}}) = 1$ then $\operatorname{Span}(\widehat{\boldsymbol{B}}) \equiv \operatorname{Span}(\widetilde{\boldsymbol{B}})$, with decreasing values which indicate increasing distances; if $R^2(\widetilde{\boldsymbol{B}}) = 0$ then the two subspaces are orthogonal.

A Criterion for Selecting a Subset of Variables

Since $R^2(\widetilde{\boldsymbol{B}})$ is simply maximized when $\widetilde{\boldsymbol{B}} \equiv \widehat{\boldsymbol{B}}$, i.e. no coefficients are set to zero, a reasonable criterion must include a penalty term which penalizes larger subsets. For a subset of predictors $\mathscr{S}_k \subset \{1, 2, \ldots, p\}$ of $\dim(\mathscr{S}_k) = k$, a sparse basis $\widetilde{\boldsymbol{B}}_k$ is obtained by setting to zero the coefficients for the $(p-k)$ predictors not included in $\mathscr{S}_k$. The criterion adopted, as motivated by Chipman and Gu (2005), is based on the maximization of the following convex linear combination:

$$C_k = (1-w)\, R^2(\widetilde{\boldsymbol{B}}_k) + w\, \frac{p-k}{p} \propto (1-w)\, R^2(\widetilde{\boldsymbol{B}}_k) - w\, \frac{k}{p} \tag{4}$$

where $w \in [0, 1]$ is a tuning parameter which controls the amount of penalty imposed. The criterion in (4) balances a measure of how well the sparse basis approximates the unconstrained estimated basis, and a penalization term which is an increasing function of the fraction k/p of predictors having nonzero coefficients among the original p predictors. Such penalty term can be seen as a form of L_0-norm penalization. If $w = 0$ then no penalty is introduced and the criterion in (4) selects $\widetilde{\boldsymbol{B}}_k \equiv \widehat{\boldsymbol{B}}$. As w increases, the solution maximizing (4) becomes more sparse. Empirical evidence suggests to adopt $w = 0.2$ as a reasonable default value in practice.

Algorithms for Subset Selection

The number of possible subsets of k variables from a total of p is given by $\binom{p}{k}$. Thus, the space of all possible subsets of size k ranging from 1 to p has number of elements equal to $\sum_{k=1}^p \binom{p}{k} = 2^p - 1$. An exhaustive search becomes unfeasible even for moderate values of p. To alleviate this problem, we may adopt a *stepwise*

search algorithm. At each stage it searches for the predictor to add which maximizes the squared trace correlation in (3) among the covariates not already selected, and then it assesses whether one predictor in the current subset could be dropped once the new predictor is included. During both the forward and the backward step, directions are orthogonalized to ensure that $\widetilde{\boldsymbol{B}}^\top \widehat{\boldsymbol{\Sigma}}\, \widehat{\boldsymbol{B}} = \boldsymbol{I}_p$, where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. These steps are iterated until all the predictors have been included. The "best" subset is then selected on the basis of the criterion in (4).

A similar technique is the *sequential replacement* search (Miller 2002). Here the basic idea is that once two or more variables have been included in the active subset, we inspect whether any of those included variables can be replaced with another one which provides a larger squared trace correlation. The sequential replacement step can be implemented on a forward search or using random starts. Again, the "best" subset is selected on the basis of the criterion in (4).

The above mentioned algorithms are not guaranteed to lead to the global optimum, in particular if p is large and the predictors are not orthogonal. Alternatively, and perhaps more efficiently, branch and bound algorithms (Hand 1981) or genetic algorithms (Goldberg 1989) could be used to directly maximize the criterion in (4). These alternative search strategies are currently under study.

3 Synthetic Data Examples

In this section we discuss the results of some simulation studies where the proposed approach is applied using two dimension reduction methods, namely SIR and SAVE, and a comparison is made with LASSO and stepwise OLS regression.

OLS is a consistent method for estimating a single direction (Li and Duan 1989), while stepwise subset selection is often applied for the purpose of variable selection, implicitly adopting a L_0-norm penalization. LASSO is a constrained version of OLS, which involves a L_1-norm penalization, thus effectively shrinking some of the coefficients toward zero. In both cases, the amount of penalization can be selected using the C_p criterion (Efron et al. 2004).

The accuracy of each method is assessed by comparing the angle formed by the estimated and the sparse bases with the true basis. The subset selection procedure is evaluated by computing both the true inclusion rate (TIR), i.e. the ratio of the number of correctly identified active predictors to the number of truly active predictors, and the false inclusion rate (FIR), i.e. the ratio of the number of falsely identified active predictors to the total number of inactive predictors. These measures are also known as sensitivity and 1-specificity, respectively, and, ideally, we wish to have TIR to be close to 1 and FIR to be close to 0 at the same time.

Model I. Consider the data generated in Sect. 1 for the simple 1D model $Y = \exp(-0.75\beta^\top X + 1) + \epsilon$, with $\beta = (1, -1, 0, \ldots, 0)$. The left panel of Fig. 1 shows the squared trace correlation between the estimated SIR basis and the sparse basis selected by the criterion in (4) as the tuning parameter is varied: values of $w \in [0.1, 0.7]$ provide stable subset selection solutions, with the first two variables which

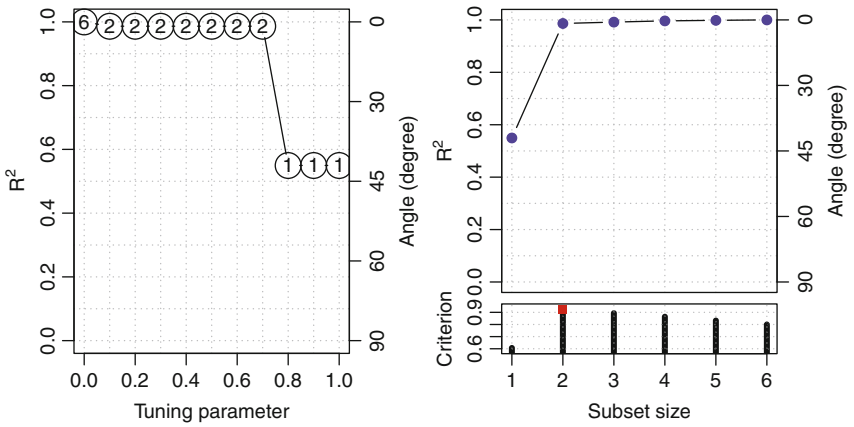


Fig. 1 Results of subset selection procedure for the simple 1D model. *Left panel*: plot of $R^2(\widetilde{\boldsymbol{B}})$ for sparse basis solutions at different values of the tuning parameter w – values within *circles* refer to subset sizes. *Right panel*: plot of $R^2(\widetilde{\boldsymbol{B}})$ at each step of the stepwise search, with a graph of the C_k criterion for $w = 0.2$ at the *bottom*

are correctly selected. However, this behavior is not to be expected in general: in most cases as w grows the size of the subsets decreases since larger subsets are increasingly penalized. Finally, note that for $w \geq 0.8$ the penalty term is dominant and only one variable is selected.

The right panel of Fig. 1 shows the trace of $R^2(\widetilde{\boldsymbol{B}}_k)$ for increasing subset sizes ($k = 1, \ldots, 6$), whereas the plot at the bottom shows the C_k criterion for $w = 0.2$, which clearly indicates that a two-variables subset (formed by the first two predictors) is needed. Thus, the first two variables selected gives $R^2(\widetilde{\boldsymbol{B}}) = 0.9864$, corresponding to an angle of 6.7° with the original SIR basis.

We conducted a Monte Carlo study by replicating the above analysis 500 times for each combination of sample sizes $n = \{100, 200, 500, 1,000\}$, and number of predictors $p = \{5, 10, 20\}$. In Table 1 we report the averages of the angle formed by the estimated and the sparse bases with the true basis of the subspace. The estimation accuracy improves for all the methods as the sample size increases, with sparse basis estimates which produce uniformly smaller angles on average. SIR appears to be the most accurate, followed by LASSO and Stepwise OLS, whereas SAVE is extremely sensible to the number of observations available when p is large.

From Table 2 we can see that SIR, LASSO, and Stepwise OLS, correctly identify the true active predictors, but SIR is able to discard the irrelevant predictors, while LASSO and, to a less extent, Stepwise OLS are also selecting inactive predictors. SAVE is less accurate on both TIR and FIR, with the previous comments which apply also in this case. Overall, SIR sparse solutions recovered the true active predictors, which allowed to improve subspace estimation accuracy.

Model II. Consider the quadratic model $Y = (\beta^\top X)^2 + 0.5\epsilon$, with $\beta = (1, -1, 0, \ldots, 0)$, where the predictors $X = (X_1, X_2, \ldots, X_p)$ follow independent standard normal distributions, and they are independent of the error component $\epsilon \sim N(0, 1)$.

Table 1 Results from the simulation study for Model I: average values of the angle (in degree) between the estimated basis with the true basis, $\angle(\widehat{\boldsymbol{B}}, \boldsymbol{B})$, and the sparse basis with the true basis, $\angle(\widetilde{\boldsymbol{B}}, \boldsymbol{B})$, as a function of sample sizes (n) and different number of predictors (p)

p	n	$\angle(\widehat{\boldsymbol{B}}, \boldsymbol{B})$			$\angle(\widetilde{\boldsymbol{B}}, \boldsymbol{B})$			
		SIR	SAVE	OLS	SIR	SAVE	LASSO	Stepwise
5	100	8.27	30.17	12.45	4.23	28.14	10.58	10.14
	200	6.20	7.16	9.61	3.10	3.62	8.18	7.92
	500	3.67	3.67	6.47	1.84	1.86	5.50	5.33
	1000	2.68	2.56	4.72	1.38	1.31	3.97	3.88
10	100	13.10	89.48	19.09	5.33	89.32	13.67	15.22
	200	8.51	81.77	14.30	2.90	81.54	9.79	11.17
	500	5.29	7.16	9.37	1.74	2.52	6.32	7.29
	1000	3.68	3.83	7.15	1.28	1.23	4.84	5.66
20	100	20.60	89.81	26.76	12.62	89.73	15.60	21.06
	200	12.65	89.89	20.51	3.79	89.82	11.00	15.67
	500	7.68	89.79	14.01	1.76	89.75	7.61	10.88
	1000	5.30	14.34	10.00	1.22	7.43	5.03	7.67

Table 2 Results from the simulation study for Model I: average values of TIR and FIR as a function of sample sizes (n) and different number of predictors (p)

p	n	TIR				FIR			
		SIR	SAVE	LASSO	Stepwise	SIR	SAVE	LASSO	Stepwise
5	100	1.00	0.96	1.00	1.00	0.00	0.13	0.36	0.15
	200	1.00	1.00	1.00	1.00	0.00	0.00	0.38	0.17
	500	1.00	1.00	1.00	1.00	0.00	0.00	0.37	0.17
	1000	1.00	1.00	1.00	1.00	0.00	0.00	0.37	0.15
10	100	1.00	0.59	1.00	1.00	0.01	0.67	0.31	0.16
	200	1.00	0.65	1.00	1.00	0.00	0.63	0.29	0.15
	500	1.00	1.00	1.00	1.00	0.00	0.00	0.30	0.15
	1000	1.00	1.00	1.00	1.00	0.00	0.00	0.30	0.16
20	100	1.00	0.54	1.00	1.00	0.05	0.66	0.24	0.16
	200	1.00	0.52	1.00	1.00	0.00	0.64	0.22	0.15
	500	1.00	0.50	1.00	1.00	0.00	0.63	0.23	0.16
	1000	1.00	1.00	1.00	1.00	0.00	0.01	0.22	0.16

Since the response function is symmetric it is known that SIR fails to recover the basis of the CDRS. However, SAVE can deal with such a situation. Table 3 shows the results from a simulation study like the one described previously. Again, sparse basis estimation allows to uniformly improve accuracy for SIR and SAVE, but not for LASSO and stepwise OLS. SAVE appears to be quite good at finding the correct direction, in particular as the sample size grows with respect to the number of available predictors.

From Table 4 we see that only SAVE is effectively able to recover the true active predictors and discard the irrelevant ones. SIR tends to select the active predictors, but too often also it includes irrelevant covariates. On the contrary, both LASSO and

Table 3 Results from the simulation study for Model II: average values of the angle (in degree) between the estimated basis with the true basis, $\angle(\widehat{\boldsymbol{B}}, \boldsymbol{B})$, and the sparse basis with the true basis, $\angle(\widetilde{\boldsymbol{B}}, \boldsymbol{B})$, as a function of sample sizes (n) and different number of predictors (p)

p	n	$\angle(\widehat{\boldsymbol{B}}, \boldsymbol{B})$			$\angle(\widetilde{\boldsymbol{B}}, \boldsymbol{B})$			
		SIR	SAVE	OLS	SIR	SAVE	LASSO	Stepwise
5	100	70.86	24.03	65.36	70.72	19.85	72.06	69.69
	200	78.33	11.95	64.55	78.21	6.41	71.32	68.89
	500	67.71	8.98	65.40	67.56	4.36	73.37	70.62
	1000	79.33	4.96	64.94	79.22	2.48	71.91	68.60
10	100	83.86	44.36	75.24	83.59	42.08	77.86	75.05
	200	78.36	26.54	74.12	78.19	21.03	76.10	74.38
	500	77.00	14.53	74.84	76.71	6.54	76.78	74.94
	1000	77.65	9.52	74.15	77.47	3.36	76.85	74.41
20	100	85.98	77.78	80.01	85.89	77.53	79.62	78.31
	200	86.03	52.30	80.31	85.99	50.46	80.09	78.89
	500	85.87	21.97	80.22	85.75	13.73	80.12	78.50
	1000	83.69	15.41	79.84	83.57	6.13	79.93	78.48

Table 4 Results from the simulation study for Model II: average values of TIR and FIR as a function of sample sizes (n) and different number of predictors (p)

p	n	TIR				FIR			
		SIR	SAVE	LASSO	Stepwise	SIR	SAVE	LASSO	Stepwise
5	100	0.734	0.997	0.411	0.418	0.6067	0.1293	0.2167	0.1600
	200	0.664	1.000	0.426	0.431	0.6593	0.0020	0.1980	0.1560
	500	0.755	1.000	0.382	0.400	0.5733	0.0000	0.1953	0.1493
	1000	0.671	1.000	0.423	0.439	0.6753	0.0000	0.2073	0.1520
10	100	0.659	0.968	0.344	0.386	0.6400	0.3345	0.1638	0.1562
	200	0.714	0.998	0.376	0.402	0.6110	0.1328	0.1790	0.1640
	500	0.722	1.000	0.367	0.402	0.6010	0.0102	0.1678	0.1727
	1000	0.716	1.000	0.373	0.414	0.5968	0.0000	0.1720	0.1685
20	100	0.658	0.790	0.316	0.414	0.6389	0.5989	0.1173	0.1597
	200	0.667	0.963	0.312	0.403	0.6334	0.4138	0.1291	0.1641
	500	0.649	1.000	0.313	0.411	0.6254	0.0660	0.1252	0.1593
	1000	0.701	1.000	0.341	0.421	0.6113	0.0112	0.1170	0.1523

stepwise OLS appear to discard the irrelevant predictors but also the true active predictors. These drawbacks are responsible for the worst accuracy achieved by the other methods in comparison to SAVE.

4 Conclusions

Dimension reduction methods play an important role in multivariate statistical analysis. Some of them can be seen as a linear mapping from the original feature space to a dimension reduction subspace with the aim of retaining most of the relevant

statistical information available in the data. However, some features may provide redundant information, whereas some other features may be irrelevant.

In this paper we discussed a sparse estimation approach for subset selection in dimension reduction methods based on a criterion which penalizes larger subsets. We show through simulations that the proposed approach allows to improve estimation accuracy, and it provides stable and interpretable solutions.

Acknowledgements Financial support from the project "Multinational firms in the service industry and economic performance in manufacturing", funded by the Italian Ministry of University and Scientific Research, is gratefully acknowledged.

References

Chen, C. H., & Li, K. C. (1998). Can be SIR as popular as multiple linear regression? *Statistica Sinica*, *8*, 289–316.

Chipman, A. H., & Gu, H. (2005). Interpretable dimension reduction. *Journal of Applied Statistics*, *32*(9), 969–987.

Cook, R. D. (1998a). *Regression graphics: Ideas for studying regressions through graphics*. New York: Wiley.

Cook, R. D. (1998b). Principal hessian directions revisited. *Journal of the American Statistical Association*, *93*(441), 84–94.

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics*, *32*(3), 1062–1092.

Cook, R. D., & Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association*, *100*(470), 410–428.

Cook, R. D., & Weisberg, S. (1991). Discussion of Li (1991). *Journal of the American Statistical Association*, *86*, 328–332.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, *32*(2), 407–451.

Goldberg, D. (1989). *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley Professional, Boston, MA.

Hand, D. J. (1981). Branch and bound in statistical data analysis. *The Statistician*, *30*, 1–13.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, *86*, 316–342.

Li, K. C. (1992). On principal hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, *87*(420), 1025–1039.

Li, K. C., & Duan, N. (1989). Regression analysis under link violation. *Annals of Statistics*, *17*(3), 1009–1052.

Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika*, *94*(3), 603–613.

Li, L., & Nachtsheim, C. J. (2006). Sparse sliced inverse regression. *Technometrics*, *48*(4), 503–510.

Miller, A. (2002). *Subset selection in regression* (2nd ed.). Chapman and Hall/CRC, Boca Raton.

Ni, L., Cook, R. D., & Tsai, C. L. (2005). A note on shrinkage sliced inverse regression. *Biometrika*, *92*(1), 242–247.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B-Methodological*, *58*(1), 267–288.

Local Statistical Models for Variables Selection

Silvia Figini

Abstract The objective of this paper is to find the most frequent itemsets in a database, made up of categorical explanatory variables and a continuous response variable. To achieve this aim we propose to extend local data mining techniques based on association rules. We assess the performance of our model by developing appropriate model indicators derived from classical concentration measures.

1 Introduction

In general, the objective of local data mining techniques as association rules is to underline groups of items that typically occur together in a set of transactions. The relevance of association rules is measured by descriptive statistical measures, named "measures of interest", such as the support, the confidence and the lift (see e.g., Giudici et al. 2009; Hand et al. 2001, Hastie et al. 2001).

While classical association rules deals with binary variables, in this paper we propose a novel methodology aims at selecting, in an exploratory way, which of the levels of the categorical variables at hand, can be chosen to explain a continuous response variable. In order to reach this objective we have proposed and implemented a rule induction methodology, and related measures of interestingness. What developed can be applied to any data mining problems where the aim is to describe the association structure between categorical variables in a given (large) database, with the eventual aim of predicting a continuous response variable.

The paper is organized as follows: in Sect. 2 we describe our methodological proposal; Sect. 3 introduces assessment measures; finally, Sect. 4 presents the conclusions and highlights the possibility to employ this contribution in real applications.

2 Proposal

Typically, association rules are local unsupervised techniques (Apriori algorithm, Agrawal et al. 1995) useful to generate, starting from binary and categorical variables, statistical measures of interest such as: support, confidence and lift.

S. Ingrassia et al. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-11363-5_65,

Table 1 Example data

	V_1	V_2	V_3	V_4	V_5	$\boldsymbol{T}$
	1	2	2	1	1	100
	1	2	2	2	3	200
	2	1	3	1	1	300
	1	2	3	1	2	400
	3	2	2	1	3	500
	1	2	1	1	3	600
	2	2	3	1	2	700
	2	2	3	1	1	800
	3	1	2	2	3	900
	1	2	2	2	1	1000
	2	1	1	1	3	1100
Q_3	3	1	1	2	2	1200
	3	1	2	1	3	1300
	3	1	3	2	3	1400
	2	2	3	2	2	1500
	3	1	2	2	1	1600

Table 2 Lower, global and upper relative frequencies of the modes of the categorical variables: example

			V_1	V_2	V_3	V_4	V_5
		Mode	3	2	2	1	3
global data	$[Min(T), Max(T)]$	f	37.5%	56.3%	43.8%	56.3%	43.8%
"lower" part	$[Min(T), Q_3)$	f	25.0%	66.7%	41.7%	66.7%	41.7%
"upper" part	$[Q_3, Max(T)]$	f	75.0%	25.0%	50.0%	25.0%	50.0%

The novelty of this paper is to extend association rules to a different environment. In order to describe our proposal, a quantitative response variable, T (e.g., the total profit of a shop) and a set of categorical explanatory variables are required (e.g., the questionnaire item variables). The method employs categorical variables in order to identify groups of observations highly related with the response variable.

For sake of clarity we refer to a theoretical example, reported in Table 1. Table 1 shows a dataset with 16 observations and five categorical variables ($V_1, \ldots, V_5$) corresponding to five questions derived from a questionnaire. We sort in an increasing order the whole data set with respect to the quantitative response variable (T = Target), as reported in Table 1.

For each categorical variable we compute the mode, m, based on all data, and its relative frequency (expressed in percentage), f. For example, based on Table 1, the mode of the variable V_1 corresponds to level 3, the mode of the variable V_2 corresponds to level 2, and so on.

We split the data in two parts: one "lower" part from Min(T), the minimum value of T, to Q_3, the third quartile, and one "upper" part from Q_3 to Max (T), the maximum value of T. We then consider the frequency distribution for the qualitative variables in each of the two parts. This is illustrated in Table 2, that also reports the same statistics for the whole dataset ("global").

Table 3 Lower, global and upper frequencies of the modes of the categorical variables: generalisation

			V_1	V_2	...	V_I
		Mode	m_1	m_2	...	m_I
global data	$[Min(T), Max(T)]$	f	f_1	f_2	...	f_I
"lower" part	$[Min(T), Q_3)$	f	l_1	l_2	...	l_I
"upper" part	$[Q_3, Max(T)]$	f	u_1	u_2	...	u_I

Note that in Table 2 we have derived the relative frequency of the mode in each of the two parts. For example, in the "lower" part the relative frequency of the mode of V_1 is 25% (3/12).

We remark that in this contribution we have selected Q_1 and Q_3 as quantiles; however Q_1 and Q_3 can be replaced using, for particular application, different statistical quantities.

Since the mode is not monotone, in this paper the algorithm proposed works on categorical unimodal variables.

What discussed can be generalised as reported in Table 3.

In Table 3, for $i = 1, \ldots, j, \ldots, I$ m_i is the mode of the i-th qualitative variable; l_i is the frequency of the mode in the "lower" part of the data, u_i is the frequency of the mode in the "upper" part and, finally, f_i is the frequency of the mode using the whole dataset.

To select the most influent variables, firstly, we check whether the relative frequencies of the mode values l_i, f_i and u_i are strictly monotone (e.g., in strictly increasing or decreasing order). In other words, we want to verify whether either of the following is true:

$$\begin{aligned} l_i &> u_i, \\ &or \\ l_i &< u_i. \end{aligned} \tag{1}$$

Note that when a single split occurs the condition in (1) can be restated as $l > u$ or $l < u$ without loss of generality. We call this "`monotonicity`" condition. The rationale of this condition is that we want to check whether the mode of each categorical variable is associated with the tails of the response distribution and, therefore, the categorical variable affects the continuous target. For example (1) is satisfied if most of modal levels are in the "upper" part or if most modal levels are in the "lower" part. This condition can be thought as a "`concordance filter`".

We then verify whether the difference between the relative frequencies l_i and u_i, in absolute value, is greater than a common threshold c.

$$|u_i - l_i| > c. \tag{2}$$

This second condition can be interpreted as a "`concentration condition`". In this application we have selected c on the basis of expert opinions

and knowledge on the data at hand; however, c should be selected using more sophisticated statistical criterias, as discussed in Hand et al. (2009).

Note that, based on conditions (1) and (2) only V_1, V_2 and V_4 are the categorical variables that can be deemed to affect the response target variable. In particular, we note that V_1 satisfies condition (1) because all the frequencies in each group are in increasing order as $l_i = 25\%$, $f_i = 37.5\%$ and $u_i = 75\%$. Also condition (2) is satisfied, since the difference among the frequencies u_i and l_i is greater than a specific threshold, such as 20% ($u_i - l_i = 50\%$).

We define "string" a combination of the different levels of the selected variables. We want to choose the strings that "best explain" the response variable.

Let $V_1, \ldots, V_j$ be the selected variables ($j \leq I$) (in the example $j = 4$), characterized respectively by $A_a(a = 1, \ldots, a^*)$, $B_b(b = 1, \ldots, b^*)$, $C_c(c = 1, \ldots, c^*)$, $D_d(d = 1, \ldots, d^*)$, where a^*, b^*, c^*, d^* are, respectively, the maximum number of levels for variables $V_1, \ldots, V_j$. A string is then a combination of variable levels such that: $S_k = (V_1 = a \bigwedge V_2 = b \bigwedge \ldots \bigwedge V_j = d)$, for $k = 1, \ldots, K$.

For example, if $L_1, \ldots, L_I$ are the level sets of the selected variables, a string can be defined as $l_1, \ldots, l_I$, with $l_i \in L_I$, and the total number of strings becomes $K = \prod_i |L_i|, i = 1, \ldots, I$. In an actual database there will be 'K' different strings, each of them will be denoted by S_k, or k for brevity.

In principle, we may consider all possible strings, obtained using all possible level combinations of the selected variables.

We proceed by understanding which strings are the most related with the target variable. For each string we consider its frequency in the actual dataset and we cumulate the corresponding target values. This gives us an indication of the intensity (or concentration) of the target values, corresponding to the specific string considered.

More formally, we define the Relative Cumulative Frequency (RCF_k) of the target T at string k as follows:

$$RCF_k = \frac{\sum_{i=1}^{N_k} T_{ik}}{\sum_{I}^{i=1} T}, \tag{3}$$

where, for a given string k, with absolute frequency N_k, and $i = 1, \ldots, N_k$, T_{ik} is the corresponding (continuous) target value.

For each string, we match its intensity, calculated as before, with its weight, calculating OF_k, the Observation Frequency of each string in the dataset, as follows:

$$OF_k = \frac{N_k}{N}, \tag{4}$$

where N is the total number of observations available.

In order to choose the most important strings, let D_k indicate, for each string, the difference between the intensity RCF_k and the frequency OF_k. This extends to strings what applied in concentration studies (see e.g., Gastwirth 1972). D_k will

Table 4 Strings and corresponding statistics for the example data

	V_1	V_2	V_4	RCF_k	OF_k	D_k
String 1	1	1	1	0	0	0
String 2	1	1	2	0	0	0
String 3	1	2	1	8.09%	18.75%	−10.66%
String 4	1	2	2	8.82%	12.50%	−3.68%
String 5	2	1	1	10.29%	12.50%	−2.21%
String 6	2	1	2	0	0	0
String 7	2	2	1	11.03%	12.50%	−1.47%
String 8	2	2	2	11.03%	6.25%	4.78%
String 9	3	1	1	0	0	0
String 10	3	1	2	9.56%	6.25%	3.31%
String 11	3	2	1	37.50%	25.00%	12.50%
String 12	3	2	2	3.68%	6.25%	−2.57%

play the role of the main interestingness measure of a string, as the support, confidence or the lift in classical association rules.

In terms of our running example, we had found V_1, V_2 and V_4 as potentially influent variable. Possible strings are thus made up of all possible levels combination generated by such variables, as shown in Table 4, that also reports RCF_k, OF and D for each string. Note that, for this example, the total intensity is equal to $RCF_k = 13{,}600$ and the number of available observations is $N = 16$.

From Table 4, note that variable V_1 has three possible levels $(1, 2, 3)$, while variable V_2 and variable V_3 have only two possible levels $(1, 2)$: their combination generates therefore a number of strings equal to the product of all possible levels ($3 \times 2 \times 2 = 12$ strings).

Based on Table 4, we can order strings in terms of their D_k values, the higher representing better discriminatory power. For example, we may retain only those strings whose value of D_k is greater than a set threshold, p.

We can then interpret D_k as a discriminant function or, in data mining terminology, a measure of interest aimed at string selection. For example, if we take $p = 1\%$ the selected strings will be $11, 8, 10$ who explain about 58% of the target response with a frequency of about 37%. In particular, string 11 shows a D_i value equal to 3.31%, string 8 a value equal to 4.78%, string 11 a value equal to 12.50%.

In other words, we explain 58.09% of the quantitative response variable considering only 37.5% of the observations (exactly six observations).

To further assess our approach, and compare it with other methods, in the next section we will evaluate each string using concentration and heterogeneity measures.

We conclude the description of our methodology with two remarks. First we point out that, when we have variables with many levels and the sample size is small, it is possible to obtain many level combinations which do not appear in the dataset. This problem is indeed well known, for example, in the data mining literature (see e.g., Hand et al. 2001): it is widely recognised that standard local data mining techniques for transactional data (e.g., Association Rules) show the same weakness and, for this purpose, measures of interestingness are developed to filter out "nfrequent"

rules. Similarly, in our case, all strings can potentially be selected but those never observed and, more generally, those for which the concentration measure D_k is less than a specific threshold p are discarded.

Second, our method is able to explain a large proportion of response on the basis of a small set of observation. The rationale behind this is the usage of a model performance measure (the D_k statistics) which is similar to the captured response (lift) in classical predictive data mining methods. Precisely, the D_k statistics contrasts the cumulative intensity of the target with the cumulative frequency of the strings.

3 Assessment

In order to measure the importance of each string, we can further elaborate the concept behind the D_k statistics and develop concentration measures (see e.g., Gastwirth 1972).

Based on the data at hand, two extreme situations are possible from a theoretical point of view:

1. Minimum response concentration (equi-distribution): the k strings share equal quantities of the T variable.
2. Maximum response concentration: one string has the total amount of target and the other $K - 1$ have 0.

To assess concentration, we sort in increasing order the variable T and we make the hypothesis that T is a transferable variable. A reasonable relative indicator of concentration must be close to zero when the concentration is minimum, and close to one when the concentration is maximum.

The Lorenz curve (see e.g., Lorenz 1905) can be used to furnish such indicator. An index of concentration can be defined as the relationship between the area of concentration (inclusive area between the Lorenz curve and the line of perfect equality) and the area of the triangle of unitary segment, corresponding to the case of equi-distribution.

In terms of our proposed method, an ideal string shows a similar distribution of the target among the string related observations; this means that we want to minimize the within-heterogeneity between the response values of the units belonging to the same string.

We remark that, in real applications, sometimes the entire Lorenz curve is not known, and only values at certain intervals are given. In that case, the Gini coefficient of concentration can be approximated by using various techniques for interpolating the missing values of the Lorenz curve (see e.g., Gastwirth 1972).

4 Conclusions

This paper shows a new procedure to select categorical variables and combinations of their levels (strings) in a predictive context. Our method belongs to the class of local data mining techniques, and can be considered an extension of association

rules to the case of categorical, rather than binary, itemsets. We think that this methodology could be employed in a different framework, and for real applications.

We believe that the proposed methods and the related assessment measures are a good starting point to build, in an explanatory way, a statistical model able to predict a continuous response variable on the basis of categorical explanatory variables.

References

Agrawal, R., Mannila, H., & Srikant, R., Toivonen, H., & Verkamo, A. I. (1995). Fast discovery of association rules. In *Advances in knowledge discovery and data mining* (Chapter 12). AAAI/MIT Press, California, USA.

Gastwirth, J. (1972). The estimation of the Lorenz curve and gini index. *The Review of Economics and Statistics*, *54*, 306–316.

Giudici, P., & Figini, S. (2009). *Applied data mining*. John Wiley, London, UK.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. MIT Press.

Hand, D.J. & Krzanowski, W.J. (2009). *ROC curves for continuous data*, CRC/Chapman and Hall.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. Springer.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, *9*, 209–219.

Index