Contributions to Statistics

Germán Aneiros
Ivana Horová
Marie Hušková
Philippe Vieu *Editors*

Functional and High-Dimensional Statistics and Related Fields

Springer

Contributions to Statistics

The series **Contributions to Statistics** contains publications in theoretical and applied statistics, including for example applications in medical statistics, biometrics, econometrics and computational statistics. These publications are primarily monographs and multiple author works containing new research results, but conference and congress reports are also considered.
Apart from the contribution to scientific progress presented, it is a notable characteristic of the series that publishing time is very short, permitting authors and editors to present their results without delay.

More information about this series at http://www.springer.com/series/2912

Germán Aneiros • Ivana Horová
Marie Hušková • Philippe Vieu
Editors

Functional and High-Dimensional Statistics and Related Fields

Springer

Editors

Germán Aneiros
Department of Mathematics
University of A Coruña
A Coruña, Spain

Marie Hušková
Department of Probability
and Mathematical Statistics
Charles University
Prague, Czech Republic

Ivana Horová
Department of Mathematics
and Statistics
Masaryk University
Brno, Czech Republic

Philippe Vieu
Toulouse Mathematics Institute
Paul Sabatier University - Toulouse III
Toulouse, France

ISSN 1431-1968
Contributions to Statistics
ISBN 978-3-030-47755-4 ISBN 978-3-030-47756-1 (eBook)
https://doi.org/10.1007/978-3-030-47756-1

Mathematics Subject Classification (2020): 62-07, 62G05, 62G09, 62G20, 62F12, 62F15, 62F40, 62H25

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

During the last twelve years, the International Workshop on Functional and Operatorial Statistics has become a prominent platform for exchange of ideas and communication in the growing community of researchers in functional data analysis. Following the success of the previous meetings held in Toulouse (France, 2008), Santander (Spain, 2011), Stresa (Italy, 2014) and A Coruña (Spain, 2017), the 5th IWFOS takes place at Masaryk University in Brno, Czech Republic. The workshop was originally planned for June 2020 but due to the rapidly evolving coronavirus pandemic it has been postponed. Nevertheless, this collection of peer-reviewed short papers is published as planned. It reflects the diversity of theoretical, methodological and applied advances in functional data analysis and its intersection with other areas of statistics, such as high-dimensional data analysis and nonparametric statistics, as well as the diversity of the community itself.

We would like to thank all the authors presenting their work at the workshop. We are particularly grateful to invited speakers Gérard Biau (Sorbonne Université, France), Eduardo García Portugués (Universidad Carlos III de Madrid, Spain), Lajos Horváth (University of Utah, USA), Roberto Imbuzeiro (Instituto Nacional de Matemática Pura e Aplicada, Brazil), Dominik Liebl (Rheinische Friedrich-Wilhelms-Universität Bonn, Germany), Regina Y. Liu (Rutgers School of Arts and Sciences, USA), Stanislav Nagy (Charles University, Prague, Czech Republic), Piercesare Secchi (Politecnico di Milano, Italy) and Yoav Zemel (University of Cambridge, United Kingdom) .

We especially appreciate the effort of the members of the Scientific Committee, namely John Aston (Cambridge, UK), Ricardo Cao (A Coruña, Spain), Antonio Cuevas (Madrid, Spain), Aurore Delaigle (Melbourne, Australia), Manuel Febrero (Santiago de Compostela, Spain), Ricardo Fraiman (Montevideo, Uruguay), Aldo Goia (Novara, Italy), Daniel Hlubinka (Prague, Czech Republic), Siegfried Hörmann (Graz, Austria), David Kraus (Brno, Czech Republic), Sara Lopez-Pintado (New York, USA), Steve Marron (Chapel Hill, USA), Alexander Meister (Rostock, Germany), Victor Panaretos (Lausanne, Switzerland), Greg Rice (Waterloo, Canada) and Simone Vantini (Milan, Italy), and other experts during the preparation of the scientific program and the review process.

We are grateful to the following academic and private institutions and organizations for their support: Faculty of Science, Masaryk University, Brno, Faculty of Mathematics and Physics, Charles University, Prague, Union of Czech Mathematicians and Physicists, Brno branch, Institut de Mathématiques de Toulouse, Autocont, Prefa Brno, Kiwi.com, Trilobyte Statistical Software, SC&C Partner and Home Credit.

The preparation of IWFOS is possible thanks to the members of the Organizing Committee, in particular, Enea Bongiorno, Marie Budíková, Jitka Forejtová Zhořová, Jan Koláček, Zdeněk Pospíšil, Lenka Přibylová, Petr Vitík and Jan Vondra. The preparation of the proceedings went smoothly thanks to the dedicated work of Ondřej Pokora, Jiří Zelinka and David Kraus from the Organizing Committee, and Veronika Rosteck and Gerlinde Schuster of Springer.

Brno, March 2020

Germán Aneiros
Ivana Horová
Marie Hušková
Philippe Vieu

Contents

List of Contributors

M. Carmen Aguilera-Morillo
Universitat Politècnica de València and uc3m-Santander Big Data Institute, Spain
e-mail: mdagumor@eio.upv.es

Mohamed Alahiane
Université Cadi Ayyad, Ecole Nationale des Sciences Appliquées,
Marrakech, Morocco
e-mail: alahianemed@gmail.com

Javier Álvarez-Liébana
Department of Statistics and Operations Research and Mathematics Didactics,
University of Oviedo, C/ Federico García Lorca, 18, 33007 Oviedo, Spain
e-mail: alvarezljavier@uniovi.es

Gonzalo Álvarez-Pérez
Department of Physics, University of Oviedo, C/ Federico García Lorca, 18, 33007
Oviedo, Spain
e-mail: gonzaloalvarez@uniovi.es

Germán Aneiros
Research group MODES, CITIC, ITMATI, Departamento de Matemáticas,
Facultade de Informática, Universidade da Coruña, 15071 A Coruña, Spain
e-mail: ganeiros@udc.es

Eleonora Arnone
MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da
Vinci 32, 20133 Milano, Italy
e-mail: eleonora.arnone@polimi.it

Mara S. Bernardi
MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da
Vinci 32, 20133 Milano, Italy
e-mail: marasabina.bernardi@polimi.it

Gérard Biau
Sorbonne Université, Laboratoire de Probabilités, Statistique et Modélisation,
4 place Jussieu, 75005 Paris, France
e-mail: gerard.biau@sorbonne-universite.fr

Enea G. Bongiorno
Università del Piemonte Orientale, Via Perrone 18, 28100, Novara, Italy
e-mail: enea.bongiorno@uniupo.it

Alain Boudou
Equipe de Stat. et Proba., Institut de Mathématiques, UMR5219, Université Paul
Sabatier, 118 Route de Narbonne, F-31062 Toulouse Cedex 9, France
e-mail: boudou@math.univ-toulouse.fr

Davide Burba
MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da
Vinci, 32, 20133 Milano, Italy
e-mail: davide.burba@mail.polimi.it

Carlos Cabo
Universidad de Oviedo, Spain
e-mail: cabo.gmail@uniovi.es

Lilei Cheng
School of Mathematics, Hefei University of Technology, China
e-mail: hfutcll@163.com

Alejandro Cholaquidis
Universidad de la República, Uruguay
e-mail: acholaquidis@cmat.edu.uy

Mathieu Couplet
EDF R&D, 6 quai Watier, 78400 CHATOU, France
e-mail: mathieu.couplet@edf.fr

Laurent Delsol
Université d'Orléans, Rue de Chartres,
B.P. 6759, FR-45067 Orléans cedex 2, France
e-mail: Laurent.Delsol@univ-orleans.fr

Graciela Estévez-Pérez
Departamento de Matemáticas, Universidade da Coruña, Spain
e-mail: graci@udc.es

Adeline Fermanian
Sorbonne Université, Laboratoire de Probabilités, Statistique et Modélisation,
4 place Jussieu, 75005 Paris, France
e-mail: adeline.fermanian@sorbonne-universite.fr

Federico Ferraccioli
Dipartimento di Scienze Statistiche, Via Cesare Battisti, 241, 35121 Padova, Italy
e-mail: ferraccioli@stat.unipd.it

Livio Finos
Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Via Venezia, 8, 35131 Padova, Italy
e-mail: livio.finos@unipd.it

Matteo Fontana
MOX - Department of Mathematics, Politecnico di Milano, Italy
e-mail: matteo.fontana@polimi.it

Lara Fontanella
University of Chieti-Pescara, Pescara, Italy
e-mail: lfontan@unich.it

Sara Fontanella
University of Torino, Torino, Italy and Imperial College London, UK
e-mail: s.fontanella@imperial.ac.uk,

Ricardo Fraiman
Universidad de la República, Uruguay
e-mail: rfraiman@cmat.edu.uy

Alexander Gammerman
Computer Learning Research Center - Department of Computer Science, Royal Holloway, University of London, UK
e-mail: a.gammerman@rhul.ac.uk

Yuan Gao
the Australian National University, 26C Kingsley St. ACT, Australia
e-mail: yuan.gao@anu.edu.au

Eduardo García-Portugués
Department of Statistics and UC3M-Santander Big Data Institute, Carlos III University of Madrid, Avda. Universidad 30, 28911 Leganés, Spain
e-mail: edgarcia@est-econ.uc3m.es

Aldo Goia
Università del Piemonte Orientale, Via Perrone 18, 28100, Novara, Italia
e-mail: aldo.goia@uniupo.it

Wenceslao González-Manteiga
Department of Statistics, Mathematical Analysis and Optimization, University of Santiago de Compostela, Rúa Lope Gómez de Marzoa s/n, 15782 Santiago de Compostela, Spain
e-mail: wenceslao.gonzalez@usc.es

Minh Hà Quang
RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, 15F,
Chuo-kuo, Tokyo, 103-0027, Japan
e-mail: minh.haquang@riken.jp

Harold A. Hernández-Roig
Universidad Carlos III de Madrid and uc3m-Santander Big Data Institute, Spain
e-mail: haroldantonio.hernandez@uc3m.es

Zdeněk Hlávka
Univerzita Karlova, Dept. of Probability and Mathematical Statistics, Faculty of
Mathematics and Physics, Sokolovská 83, Praha 8, Czech Republic
e-mail: hlavka@karlin.mff.cuni.cz

Daniel Hlubinka
Univerzita Karlova, Dept. of Probability and Mathematical Statistics, Faculty of
Mathematics and Physics, Sokolovská 83, Praha 8, Czech Republic
e-mail: hlubinka@karlin.mff.cuni.cz

Ivana Horová
Faculty of Science, Department of Mathematics and Statistics, Masaryk University,
Brno, Czech Republic
e-mail: horova@math.muni.cz

Lajos Horváth
University of Utah, Department of Mathematics, Salt Lake City UT 84112, U.S.A.
e-mail: horvath@math.utah.edu

Marie Hušková
Faculty of Mathematics and Physics, Charles University, Praha, Czech Republic
e-mail: huskova@karlin.mff.cuni.cz

Francesca Ieva
MOX, Department of Mathematics, Politecnico di Milano, P.zza Leonardo da Vinci
32, 20133 Milano (IT) & CADS - Center for Analysis, Decision and Society,
Human Technopole, Milano, Italy
e-mail: francesca.ieva@polimi.it

Rosaria Ignaccolo
University of Torino, Torino, Italy
e-mail: rosaria.ignaccolo@unito.it

Luigi Ippoliti
University of Chieti-Pescara, Pescara, Italy
e-mail: ippoliti@unich.it

Jan Kalina
The Czech Academy of Sciences, Institute of Computer Science, Pod Vodárenskou
věží 2, 182 07 Praha 8, Czech Republic
e-mail: kalina@cs.cas.cz

Alois Kneip
Universität Bonn, 53012 Bonn, Germany
e-mail: akneip@uni-bonn.de

Dominik Liebl
Institute of Finance and Statistics and Hausdorff Center for Mathematics,
University of Bonn, Adenauerallee 24-26, 53113 Bonn, Germany
e-mail: dliebl@uni-bonn.de

Rosa E. Lillo
Universidad Carlos III de Madrid and uc3m-Santander Big Data Institute, Spain
e-mail: rosaelvira.lillo@uc3m.es

Nengxiang Ling
School of Mathematics, Hefei University of Technology, China
e-mail: hfut.lnx@163.com

Nathalie Marie
CEA Cadarache, 13108 Saint-Paul-lez-Durance, France
e-mail: nathalie.marie@cea.fr

Amandine Marrel
CEA Cadarache, 13108 Saint-Paul-lez-Durance, France
e-mail: amandine.marrel@cea.fr

Alessandra Menafoglio
MOX - Department of Mathematics, Politecnico di Milano, Milano, Italy
e-mail: alessandra.menafoglio@polimi.it

Elsa Merle-Lucotte
LPSC Grenoble, 53 avenue des Martyrs, France
e-mail: merle@lpsc.in2p3.fr

Leonardo Moreno
Universidad de la República, Uruguay
e-mail: mrleo@iesta.edu.uy

Stanislav Nagy
Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic
e-mail: nagy@karlin.mff.cuni.cz

Fabio Nobile
CSQI - École polytechnique fédérale de Lausanne, Route Cantonale, 1015
Lausanne, Switzerland
e-mail: fabio.nobile@epfl.ch

Silvia Novo
MODES research group, CITIC, Universidade da Coruña, A Coruña, Spain
e-mail: s.novo@udc.es

Niels Lundtorp Olsen
University of Copenhagen, Denmark
e-mail: niels.olsen@math.ku.dk

Celestino Ordóñez
Universidad de Oviedo, Spain
e-mail: ordonezcelestino@uniovi.es

Idir Ouassou
Université Cadi Ayyad, Ecole Nationale des Sciences Appliquées, Marrakech and Université Mohammed VI Polytechnique, 43140 Ben Guerir, Morocco
e-mail: i.ouassou@uca.ma

Manuel Oviedo de la Fuente
Universidade de Santiago de Compostela, Spain
e-mail: manuel.oviedo@usc.es

Vic Patrangenaru
Florida State University, U.S.A.
e-mail: vpatrangenaru@fsu.edu

Davide Pigoli
Department of Mathematics, King's College London, London, United Kingdom
e-mail: davide.pigoli@kcl.ac.uk

Alessia Pini
Università Cattolica del Sacro Cuore, Italy
e-mail: alessia.pini@unicatt.it

Mustapha Rachdi
Univ. Grenoble Alpes, AGEIS laboratory, UFR SHS, BP. 47, 38040 Grenoble Cedex 09, France
e-mail: mustapha.rachdi@univ-grenoble-alpes.fr

Matthew Reimherr
Department of Statistics, Penn State University, 411 Thomas Building University Park, PA 16802, U.S.A.
e-mail: mreimherr@psu.edu

Philip T. Reiss
Department of Statistics, University of Haifa, Haifa 31905, Israel
e-mail: reiss@stat.haifa.ac.il

Javier Roca-Pardiñas
Universidade de Vigo, Spain
e-mail: roca@uvigo.es

Álvaro Rollón de Pinedo
EDF R&D, 6 quai Watier, 78400 Chatou, France
e-mail: alvaro.rollon-de-pinedo@edf.fr

Laura M. Sangalli
MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy
e-mail: laura.sangalli@polimi.it

Piercesare Secchi
MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy
e-mail: piercesare.secchi@polimi.it

Han Lin Shang
the Australian National University, 26C Kingsley St. ACT, Australia
e-mail: hanlin.shang@anu.edu.au

Chen Shen
Florida State University, U.S.A.
e-mail: cs15j@my.fsu.edu

Marta Spreafico
MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, Italy
e-mail: marta.spreafico@polimi.it

Roman Sueur
EDF R&D, 6 quai Watier, 78400 CHATOU, France
e-mail: roman.sueur@edf.fr

Massimo Tavoni
Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Italy
RFF-CMCC European Institute on Economics and the Environment (EIEE), Fondazione CMCC, Lecce, Italy
e-mail: massimo.tavoni@polimi.it

Pasquale Valentini
University of Chieti-Pescara, Pescara, Italy
e-mail: pvalent@unich.it

Simone Vantini
MOX - Department of Mathematics, Politecnico di Milano, Italy
e-mail: simone.vantini@polimi.it

Petra Vidnerová
The Czech Academy of Sciences, Institute of Computer Science, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic
e-mail: petra@cs.cas.cz

Philippe Vieu
Institut de Mathématiques, Université Paul Sabatier, Toulouse, France
e-mail: philippe.vieu@math.univ-toulouse.fr

Sylvie Viguier-Pla
Equipe de Stat. et Proba., Institut de Mathématiques, UMR5219, Université Paul Sabatier, 118 Route de Narbonne, F-31062 Toulouse Cedex 9, France,
Université de Perpignan via Domitia, LAMPS, 52 av. Paul Alduy, 66860 Perpignan Cedex 9, France
e-mail: viguier@univ-perp.fr

Meng Xu
Department of Statistics, University of Haifa, Haifa 31905, Israel
e-mail: mxu@campus.haifa.ac.il

Yanrong Yang
the Australian National University, 26C Kingsley St. ACT, Australia
e-mail: yanrong.yang@anu.edu.au

Chapter 1
An introduction to the (postponed) 5th edition of the International Workshop on Functional and Operatorial Statistics

Germán Aneiros, Ivana Horová, Marie Hušková and Philippe Vieu

Abstract This volume is composed by a set of short papers corresponding to some of the contributions that were sent to be presented at the fifth edition of the International Workshop on Functional and Operatorial Statistics (IWFOS). This fifth edition was to be held in June 2020 in Brno (Czech Republic), but had to be postponed as a consequence of the health crisis caused by the COVID-19 pandemic. The aim of this introduction is to make a fast presentation of these contributions by putting them into the recent trends in Functional Data Analysis and related fields.

1.1 IWFOS and Functional Data Analysis

The meetings IWFOS has played a major role along the last twelve years to promote Functional Data Analysis (FDA) ideas. The first edition took place in Toulouse, France (June 2008), at a moment when FDA ideas were not so much developed as they can be nowadays. Then this meeting took place each three years (Santander, Spain, 2011; Stresa, Italy, 2014; A Coruña, Spain, 2017), and each issue was the opportunity for active researchers in the field to share their recent advances and to start new collaborations. During these twelve years, all the leaders in research on FDA have participate in some way in these events (either as member of the

Germán Aneiros (✉)
Research group MODES, CITIC, ITMATI, Departamento de Matemáticas, Facultade de Informática, Universidade da Coruña, 15071 A Coruña, Spain, e-mail: ganeiros@udc.es

Ivana Horová
Faculty of Science, Department of Mathematics and Statistics, Masaryk University, Brno, Czech Republic, e-mail: horova@math.muni.cz

Marie Hušková
Faculty of Mathematics and Physics, Charles University, Praha, Czech Republic, e-mail: huskova@karlin.mff.cuni.cz

Philippe Vieu
Institut de Mathématiques, Université Paul Sabatier, Toulouse, France, e-mail: philippe.vieu@math.univ-toulouse.fr

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_1

Programs Committee or as contributor, or both of them). The fifth edition of IWFOS was to be held in June 2020 in Brno (Czech Republic), but had to be postponed as a consequence of the health crisis caused by the COVID-19 pandemic. As can be noted from the short papers included in this volume (which had been accepted for such fifth edition), the IWFOS planned for 2020 maintained the high quality standards of the other four past editions.

From a methodological point of view, one can say that FDA's ideas have been of influence on almost all the fields of Statistics, including: linear, semiparametric and nonparametric modelling, as well as regression, clustering and classification problems, or independent, time series and spatial datasets, ... Also, from an applied point of view, FDA's ideas have been used to analyse scopes of real data coming from most of applied scientific fields, including medicine, econometrics, environmetrics, physics, spectrometry, and many other ones ... This wide degree of interest of FDA's ideas is attested by the recent bibliographical studies (see for instance, [3], [4], [6], [7], [5], [1], [2], ...).

Since the third edition in 2014, and because of the wide set of links existing between FDA and High-Dimensional Statistics (HDS), the topics of IWFOS meetings have been extended to HDS (see [4] and [1]). The (postponed) 2020's edition had followed this opening strategy, that has been nicely appreciated by the participants of past editions, and extended it as well to HDS as to other related fields.

1.2 Presentation of the various chapters

This volume contains contributions on several topics in functional and high-dimensional Statistics and related fields, including:

- Classification: see Chapter **28** for a proposal, based on distance correlation, for selecting optimum scales for supervised classification of 3D point clouds.
- Confidence bands: see Chapter **21**, which focuses on a new approach, based on random field theory, for constructing simultaneous confidence bands in the case of the function-on-scalar linear regression model.
- Density estimation: see Chapter **11** for a proposal of a nonparametric method for density estimation over two-dimensional domains.
- Depth: see Chapter **25**, where the depth in finite-dimensional spaces is introduced, and it is outlined particular difficulties one faces when attempting to generalize depths to the situation of functional or other infinite-dimensional data.
- Diagnostic tests: see Chapter **10**, where a diagnostic test is constructed by using a novel procedure that allows to indicate if one functional data precedes to another one.
- Dimension reduction: see Chapter **6** for a reconstitution, based on PCA, of a cyclostationary random function; and Chapter **19** for FPCA combined with the survival Cox regression model.

- Estimation on manifolds: see Chapter 7 for some asymptotic properties related to an estimator of the level sets of a density; and Chapter 16 for stringing via manifold learning.
- High- and infinite-dimensional Statistics: see Chapter 26 for an algorithm to select impact points in a new sparse semiparametric functional model.
- Inference on functional data: see Chapter 18 for discussion related to some results (very useful in econometrics) on functional data whose mean and covariance are expanded in certain particular basis.
- Networks: see Chapter 13 for an approach to extend network analytical tools to the functional data setting; and Chapter 20 for robust neural networks.
- Operatorial Statistics: see Chapter 24 for distances between covariance operators associated with functional random processes.
- Prediction: see Chapter 12 for an algorithm to generate nonparametric prediction bands for a functional-on-scalar linear regression model.
- Regression: see Chapter 8, where the behaviour of a cross-validation approach to select the pseudo-metric is studied by means a simulation study; Chapter 22 for estimation of the functional single index regression model with responses missing at random for strong mixing time series data; and Chapter 29 for rates of convergence and asymptotic distribution of estimators in generalized functional partially linear single-index models.
- Robustness: see Chapter 30 for functional outlier detection.
- Sequential learning: see Chapter 4, where a novel signature approach is discussed, focusing in its use in machine learning.
- Small-ball probability: see Chapter 5 for an overview on asymptotic results related to a factorization of the small-ball probability, as well as illustrations of new results.
- Smoothing: see Chapter 14 for a proposal to retrieve functional data from the corresponding observed discretized valued, considering a factor model on the measurement error term.
- Spatial data: see Chapter 2 for an application of space-time regression; Chapter 3 for a simulation study related to spatial regression with partial differential equation regularization; and Chapter 23 for an overview, including application, on object oriented spatial Statistics focused on the problem of kriging prediction.
- Testing: see Chapter 9, where a test procedure for checking the validity of the single functional index model is introduced and its performance is analyzed by means of Monte Carlo experiments; Chapter 15 for a goodness-of-fit test for the functional linear model with functional response, the corresponding statistics being calibrated through a wild bootstrap on the residuals; Chapter 17 for two-sample tests based on empirical characteristic functionals; Chapter 27 for local inference controlling the false discovery rate; and Chapter 32 for adjusted p-values based on envelope tests.
- Topological object data analysis: see Chapter 31, where it is presented methodology to study distributions on object spaces.

Finally, it is worth being noted that some of these chapters include, in addition to methodology and/or asymptotics and/or simulation studies and/or overviews on some topic, interesting applications to real data, concerning the areas of:

- Automobile engineering: see Chapter **20**.
- Criminology: see Chapter **20**.
- Drawing recognition: see Chapter **4**.
- Econometrics: see Chapters **20** and **21**.
- Environmetrics: see Chapters **12**, **13**, **15**, **23** and **27**.
- Medicine: see Chapters **19**, **31** and **32**.
- Mobile phone: see Chapter **2**.
- Spectrometrics: see Chapter **26**.
- Urban environment: see Chapter **28**.

References

[1] Aneiros, G., Cao, R., Fraiman, R., Genest, C., Vieu, P.: Advances in functional data analysis and high-dimensional statistics. J. Multivariate Anal. **170**, 1–9 (2019)

[2] Aneiros, G., Cao, R., Vieu, P.: Editorial on functional data analysis and related topics. Comput. Statist. **34**, 447–450 (2019)

[3] Cuevas, A.: A partial overview of the theory of statistics with functional data. J. Statist. Plann. Inf. **147**, 1–23 (2014)

[4] Goia, A., Vieu, P.: An introduction to recent advances in high/infinite dimensional statistics. J. Multivariate Anal. **46**, 1–6 (2016)

[5] Kokoszka, P., Reimherr, M.: Introduction to Functional Data Analysis. CRC Press (2017)

[6] Müller, H.G.: Peter Hall, functional data analysis and random objects. Ann. Statist. **44**, 1867–1887 (2016)

[7] Wang, J-L., Chiou, J-M., Müller, H.G.: Functional data analysis. Annu. Rev. Stat. Appl. **3**, 257–295 (2016)

Chapter 2
Analysis of Telecom Italia Mobile Phone Data by Space-time Regression with Differential Regularization

Eleonora Arnone, Mara S. Bernardi, Laura M. Sangalli and Piercesare Secchi

Abstract We apply spatio-temporal regression with partial differential equation regularization to the Telecom Italia mobile phone data. The technique proposed allows to include specific information on the phenomenon under study through a definition of the non-stationary anisotropy characterizing the spatial regularization based on the texture of the domain on which the data are observed.

2.1 Space-Time Regression with Differential Regularization

The analysis of functional data with spatial dependence has been of great interest in the last years and various methods have been recently proposed to deal with this kind of data [10]. In this work, we consider spatial regression methods with Partial Differential Equation (PDE) regularization [12, 13, 4, 5]. In particular, we consider the Space-Time regression with PDE penalization method (ST-PDE) introduced in [7] and extend it to deal with observations featuring complex spatial dependency.

ST-PDE is a penalized regression method that models separately the spatial and the temporal regularization by considering two roughness penalties, which account separately for the regularity of the field in space and in time by using a tensor product, following the approach used also by [1, 3, 9]; while, in the generalization

Eleonora Arnone
MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy, e-mail: eleonora.arnone@polimi.it

Mara S. Bernardi (🖂)
MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy, e-mail: marasabina.bernardi@polimi.it

Laura M. Sangalli
MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy, e-mail: laura.sangalli@polimi.it

Piercesare Secchi
MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy, e-mail: piercesare.secchi@polimi.it

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_2

of the technique proposed by [2], a single roughness penalty is used to jointly model the spatial and temporal dimensions. Therefore, in the ST-PDE model, the field is estimated minimizing a functional composed by three parts: a data-fitting part, a penalization for the spatial regularity, and a penalization for the temporal regularity. In [7], the spatial penalization involves a simple differential operator that imposes smoothness to the solution. Instead, in this work, we consider a spatial penalization involving a more general PDE, that allows to impose non-stationary anisotropy to the solution, thus modeling more complex spatial dependencies. Moreover, the PDE can model problem-specific knowledge on the phenomenon under study. For example, if the PDE governing the physical phenomenon generating the data is available, it can be exploited in the spatial regularization term of the ST-PDE functional, thus driving the estimation towards a physically sound solution. In the context of the analysis of mobile phone data, where no physical knowledge on the phenomenon under study is available, we use the PDE to include in the model information about the texture of the spatial domain; in particular, we here characterize the PDE using the road network, which highly influences the data. This application highlights the high flexibility of the definition of spatial dependence imposed by the ST-PDE model.

Section 2.2 describes the Telecom Italia mobile phone data. Section 2.3 presents the model and how the texture of the domain can be used to estimate the non-stationary anisotropy characterizing the regularization.

2.2 Telecom Italia Mobile Phone Data

We consider the Telecom Italia database, provided by Convenzione di Ricerca DiAP–Politecnico di Milano and Telecom Italia. This dataset concerns the usage of mobile phone data in the metropolitan area of Milan. It collects the measurements of the Erlang, a dimensionless unit calculated by adding up the length of all the calls made by mobile phones within a region of the spatial domain in a time interval, and dividing the sum by the length of the time interval. In the case of the Telecom Italia database, Erlang data are collected over time intervals of 15 minutes from Wednesday, March 18th 2009, 00:15 to Tuesday, March 31st 2009, 23:45 on a uniform lattice of 97×109 sites with dimension 232m×309m covering the metropolitan area of Milan. In Figure 2.1, the top panel shows the map of the metropolitan area of Milan on which the data are observed, the central panel shows the Erlang data for a fixed time instant, the bottom panel shows the data in a fixed spatial location.

Since the Erlang is a measurement of the average number of active mobile phones, these data can be considered as an approximation of the number of people present in the considered sites during the sampling time windows. Therefore, the goal of the analysis of these data is the study of the population distribution and dynamics. Indeed, this dataset has been used in the context of the Green Move Project, an interdisciplinary research project financed by Regione Lombardia and focused on the development of a vehicle sharing system. Some works on this dataset are [8, 14, 17, 11, 15].

The data can be interpreted as a sampling of temporal curves with spatial dependencies; equivalently, they can also be interpreted as a sampling of spatial surfaces

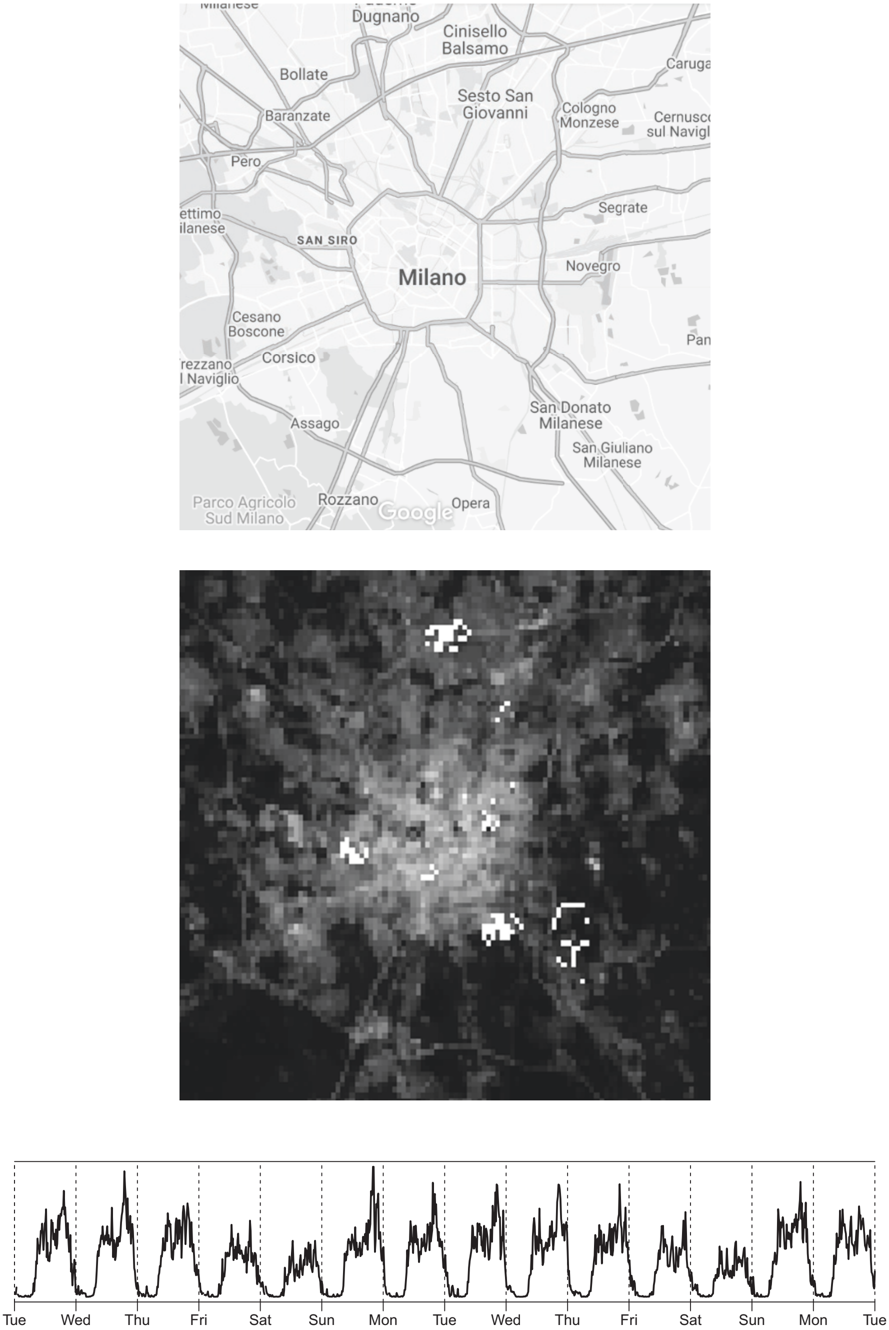


Fig. 2.1 Telecom Italia mobile phone data. Top panel: the metropolitan area of Milan, the spatial domain of the dataset. Central panel: data for a fixed time instant (white corresponds to missing data). Bottom panel: evolution in time of the data for a fixed spatial location.

with temporal dependencies. In both interpretations, the data are functional in nature and using the functional data analysis framework allows us to properly characterize the complex dependencies and extract meaningful results.

Furthermore, the data are integrals over both time and space of the quantity of interest, since each Erlang datum is a cumulative measurement over a 15-minutes time interval and a 232m×309m site. Therefore, the analysis should properly take into account the fact that the data are areal in space and integral in time.

Moreover, as Figure 2.1 shows, the spatial distribution of the data is strongly influenced by the characteristics of the urban area considered. Therefore, it is of paramount importance to take into consideration the spatial dependence driven by physical phenomenon generating the data, i.e. the population dynamics in the metropolitan area of Milan, and to adapt the estimation technique to properly take into account the characteristics of the specific urban configuration under study.

Next section deals with the characterization of the spatial dependence of the data through the definition of a penalization term involving a non-stationary anisotropic diffusion operator which represents the structure of the underlying spatial domain.

2.3 ST-PDE Model and Estimating the Non-stationary Anisotropy

In the ST-PDE functional, the classical square L^2-norm of the second derivative is employed for the temporal penalty, while we need a term which allows us to model non-stationary anisotropy for the spatial penalty. This is obtained by penalizing the misfit from a diffusion PDE $-\text{div}(K(\mathbf{p})\nabla f) = 0$, where $K(\mathbf{p})$ is a function defined on the spatial domain, taking values in the space of symmetric and positive definite 2×2-matrices. When K is a constant function equal to the identity matrix all over the spatial domain, the smoothing is isotropic in space (which is the case considered in [7]); otherwise, the smoothing is anisotropic. If, moreover, K is non-constant as a function of the spatial location $\mathbf{p}$, the smoothing is non-stationary. In our work, we exploit the texture of the spatial domain to estimate the symmetric tensor $K(\mathbf{p})$.

We can observe, from Figure 2.1, that the number of active phones presents localized strongly anisotropic features in correspondence of the main roads. Thus, we want to use the information about the morphology of the road network of the city to include non-stationary anisotropy in the ST-PDE model. The motivation for our choice is that, when we deal with cars moving on highways, we know that it is more probable that these cars will stay in the highway then that they will exit. Thus, for the spatial locations corresponding to main roads, we want to impose anisotropic smoothing that smooths more in the direction tangential to the road, and less in the other directions.

We use data form Regione Lombardia about the road network of the metropolitan area of Milan (see Figure 2.2, left panel), in order to estimate $K(\mathbf{p})$ from the city texture. In particular, we select the main roads and highways (see Figure 2.2, right panel) and exploit the orientation of the roads to identify the direction of the major axis of $K(\mathbf{p})$, i.e. the eigenvector corresponding to the larger eigenvalue. Indeed, for each spatial location **p**, the direction of the major axis of $K(\mathbf{p})$ can be defined by looking at the road map at a small scale that allows to consider one road at a time. Where no roads are present, the isotropic diffusion operator is used.

The intensity of the anisotropy can be set either exploiting prior knowledge on the phenomenon (for example, the speed limits of the roads) or extracting information from the data using an approach similar to [6], which proposes to estimate the anisotropy directly from the data.

The use of the ST-PDE model with a spatial regularization involving a non-stationary and anisotropic diffusion differential operator carrying information about the road network is particularly useful in the analysis of Telecom Italia mobile phone data, since this technique is able to suitably capture the non-trivial spatial dependencies of the observed data.

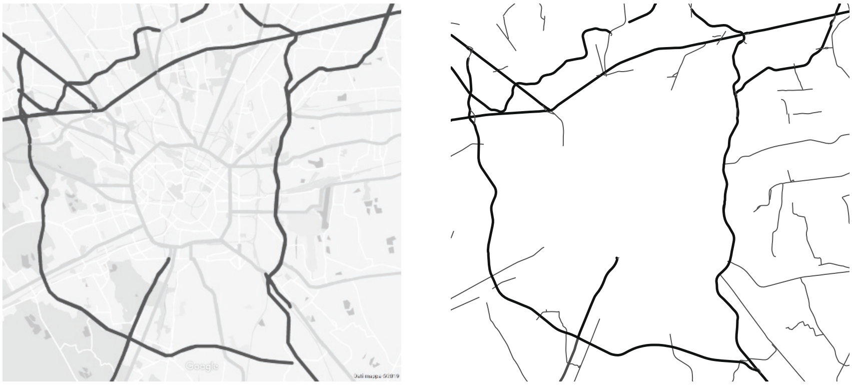


Fig. 2.2 Road network in the metropolitan area of Milan. Left panel: a view of the area from Google maps which includes main roads, secondary roads, highways and railways. Right panel: main roads and highways from www.geoportale.regione.lombardia.it used to estimate $K(\mathbf{p})$.

References

[1] Aguilera-Morillo, M.C., Durbán ,M., Aguilera, A.M.: Prediction of functional data with spatial dependence: a penalized approach. Stochastic Environ Res Risk Assess **31**, 7–22 (2017)

[2] Arnone, E., Azzimonti L., Nobile F., Sangalli L. M.: Modeling spatially dependent functional data via regression with differential regularization. Journal of Multivariate Analysis **170**, 275–295 (2019)

[3] Augustin, N.H., Trenkel, V.M., Wood, S.N., Lorance, P.: Space-time modelling of blue ling for fisheries stock management. Environmetrics **24**(2), 109–119 (2013)

[4] Azzimonti, L., Nobile, F., Sangalli, L.M., Secchi, P.: Mixed Finite Elements for Spatial Regression with PDE Penalization. SIAM/ASA Journal on Uncertainty Quantification **2**(1), 305–335 (2014)

[5] Azzimonti, L., Sangalli, L.M., Secchi, P., Domanin, M., Nobile, F.: Blood flow velocity field estimation via spatial regression with PDE penalization. Journal of the American Statistical Association **110**(511), 1057–1071 (2015)

[6] Bernardi, M.S., Carey, M., Ramsay. J.O., Sangalli, L.M.: Modeling spatial anisotropy via regression with partial differential regularization. Journal of Multivariate Analysis **167**, 15–30 (2018)

[7] Bernardi, M.S., Sangalli, L.M., Mazza, G., Ramsay, J.O.: A penalized regression model for spatial functional data with application to the analysis of the production of waste in Venice province. Stochastic Environ Res Risk Assess **31**, 23–38 (2017)

[8] Manfredini, F., Pucci, P., Secchi, P., Tagliolato, P., Vantini, S., Vitelli, V.: Treelet decomposition of mobile phone data for deriving city usage and mobility pattern in the milan urban region. In: Advances in complex data modeling and computational methods in statistics, pp. 133–147. Springer (2015)

[9] Marra, G., Miller. D.L., Zanin. L.: Modelling the spatiotemporal distribution of the incidence of resident foreign population. Statistica Neerlandica **66**(2), 133–160 (2012)

[10] Mateu, J., Romano, E.: Advances in spatial functional statistics. Stochastic Environ Res Risk Assess **31**, 1–6 (2017)

[11] Passamonti, F.: Spatio-temporal mobile phone data in Milan: Bagging-Voronoi exploration and modeling through soil use and land cover data. Master's thesis, Politecnico di Milano, MOX - Dipartimento di Matematica (2016)

[12] Ramsay, T.: Spline smoothing over difficult regions. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **54**(2), 307–319 (2002)

[13] Sangalli, L.M., Ramsay, J.O., Ramsay, T.O.: Spatial spline regression models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **75**(4), 681–703 (2013)

[14] Secchi, P., Vantini, S., Vitelli, V.: Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan. Statistical Methods Applications **24**(2), 279–300 (2015)

[15] Secchi, P., Vantini, S., Zanini, P.: Analysis of Mobile Phone Data for Deriving City Mobility Patterns. In Electric Vehicle Sharing Services for Smarter Cities, pp. 37–58. Springer, Cham (2017)

[16] Xun, X., Cao, J., Mallick, B., Maity, A., Carroll, R.J.: Parameter Estimation of Partial Differential Equation Models. Journal of the American Statistical Association **108**(503), 1009–1020 (2013)

[17] Zanini, P., Shen, H., Truong, Y.: Understanding resident mobility in Milan through independent component analysis of Telecom Italia mobile usage data. The Annals of Applied Statistics **10**(2), 812–833 (2016)

Chapter 3
Some Numerical Test on the Convergence Rates of Regression with Differential Regularization

Eleonora Arnone, Alois Kneip, Fabio Nobile and Laura M. Sangalli

Abstract We numerically study the bias and the mean square error of the estimator in Spatial Regression with Partial Differential Equation (SR-PDE) regularization. SR-PDE is a novel smoothing technique for data distributed over two-dimensional domains, which allows to incorporate prior information formalized in term of a partial differential equation. This technique also enables an accurate estimation when the shape of the domain is complex and it strongly influences the phenomenon under study.

3.1 Introduction

Spatial functional statistic is a field of research of strong interest in recent years, due to the fact that spatially dependent functional data are increasingly available in many applied fields, such as biology, life science, environmental science and engineering (see [7, 17] for a review on the recent proposed methods).

In this work, we numerically investigate the asymptotic properties of the estimator in Spatial Regression with Partial Differential Equation regularization (SR-PDE) introduced in [18, 20, 3]. SR-PDE is a penalized regression method, that includes the penalty term the misfit from a linear Partial Differential Equation (PDE). This allow a great flexibility of the method. In particular, the PDE in the regularizing

Eleonora Arnone (✉)
Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy,
e-mail: eleonora.arnone@polimi.it

Alois Kneip
Universität Bonn, 53012 Bonn, Germany, e-mail: akneip@uni-bonn.de

Fabio Nobile
CSQI - École polytechnique fédérale de Lausanne, Route Cantonale, 1015 Lausanne, Switzerland,
e-mail: fabio.nobile@epfl.ch

Laura M. Sangalli
Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy,
e-mail: laura.sangalli@polimi.it

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_3

term enables the modelling of anisotropy and non-stationarity of the phenomenon under study. Moreover, thanks to the use of the finite element method, it allows to consider domain of complex shape, such as domains with strong concavities, that affect the phenomenon under study, and to impose boundary conditions. Smoothing is a fundamental step in most analyses involving functional data [19, 10, 14]. In this respect, the considered SR-PDE method provides a versatile tool for the smoothing of functional data observed over two-dimensional domains.

Other regularized least-square smoothers have been proposed that can deal with complex domains, such as bivariate splines over triangulations [15, 11, 8, 16], soap film smoothing [24], and low-rank thin-plate spline approximations [23, 21]. All these methods have isotropic regularizing terms. Among the methods mentioned above, the only one that can comply with boundary conditions is soap film smoothing. The asymptotic properties of bivariate splines over triangulations are investigated in [16]. To the best of our knowledge, no results on large sample properties is available for any of the other methods.

The study of the asymptotic properties of classical penalized regression estimators is a well established literature that dates back to the 80s (see, e.g., [9] and references therein). The arguments used to prove the study the bias and the MSE of thin-plate-splines and of smoothing splines [4, 5, 6, 12, 13], however, exploit the existence of an explicit closed form of the Green functions of the differential operator in the regularizing term. Due to the more complex penalty considered by SR-PDE, and moreover, due to the presence of boundary conditions which enable to deal with domains of complex shape, a closed form for the Green functions of the differential operator in the regularizing term is not available for SR-PDE. In addition, as already mentioned, the estimation problem is solved by means of finite elements, with a mixed formulation. This is very convenient from a computational point of view, but makes the analysis of the asymptotic properties much more involved. In [2] a first attempt to study the bias of the infinite dimensional estimator with respect to the smoothing parameter is presented, while the finite element estimator is studied letting the discretization becomes more and more fine, but fixing the number of observations.

In this work, instead, we want to study the asymptotic behaviour of the estimator when the number of observations increases to infinity. Next section presents the estimator, while the last section reports some simulation studies that investigate the rates for the bias and the mean square error of the estimator.

3.2 Spatial Regression with PDE Penalization

Let $\Omega \subset \mathbb{R}^2$ a bounded domain, with boundary $\partial\Omega \in C^2$ or polygonal. Consider n observations $z_i \in \mathbb{R}$, for $i = 1, \ldots, n$, located at points $\mathbf{p}_i = (x_i, y_i) \in \Omega$. Assume that:

$$z_i = f_0(\mathbf{p}_i) + \varepsilon_i$$

where $f_0 : \Omega \to \mathbb{R}$ is the field we wish to estimate, and ε_i are independent errors with zero mean and finite variance σ^2.

Denote by $H^2(\Omega)$ the Sobolev space of functions in $L^2(\Omega)$ with derivatives up to the 2-th order in $L^2(\Omega)$, and let V_α the space H^2 with Dirichlet or Neuman boundary conditions, that is

$$V_\alpha = V_\alpha^{\text{dir}} = \{f \in H^2(\Omega) : \ f = \alpha \text{ on } \partial\Omega\}$$

or

$$V_\alpha = V_\alpha^{\text{neu}} = \{f \in H^2(\Omega) : \ \frac{\partial f}{\partial \nu} = \alpha \text{ on } \partial\Omega\}$$

where ν denotes the normal versor to the boundary $\partial\Omega$, and α is the value imposed on the boundary. SR-PDE solves the following estimation problem:

$$\hat{f} = \underset{f \in V_\alpha}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (f(\mathbf{p}_i) - z_i)^2 + \lambda_n \int_\Omega (Lf - u)^2 \tag{3.1}$$

where

$$L(\mathbf{p})f = -\text{div}(\mathbf{K}(\mathbf{p})\nabla f) + \mathbf{b}(\mathbf{p}) \cdot \nabla f + c(\mathbf{p})f$$

is a second order linear elliptic operator and the PDE $Lf = u$ partially describes the phenomenon under study. The smoothing parameter $\lambda_n > 0$ controls the relative weight of the two terms in the functional in (3.1): a data fidelity term, given by the sum of square errors, and a model fidelity term, the differential regularization, defined as the $L^2(\Omega)$-norm of the misfit with respect to the PDE. We explicitly highlight the dependence of the smoothing parameter with respect to n, since as the number of data locations increases less regularization is needed. We thus expect to let λ_n go to zero as n goes to infinity.

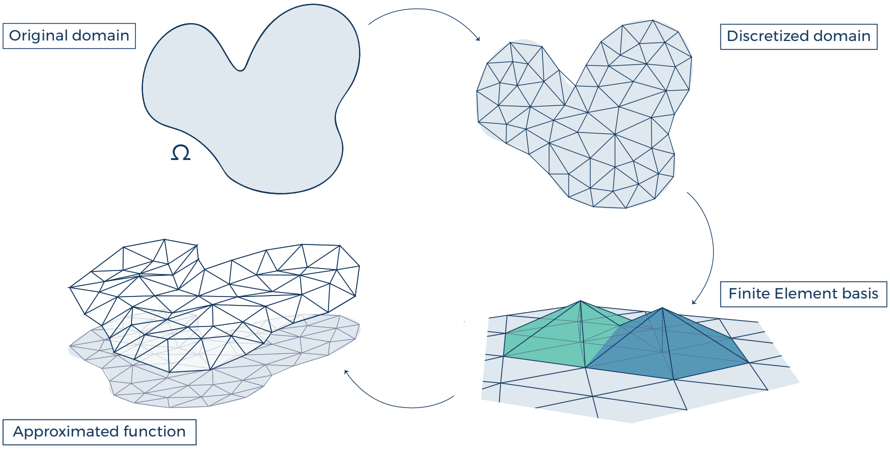


Fig. 3.1 The discretization process. Starting from the original domain (top, left), a polygonal approximation is given and a triangulation is defined (top, right). The linear finite element basis (bottom, right) is introduced over the triangulation and a piecewise linear approximation (bottom, left) of the function of interest is computed.

The SR-PDE estimator defined in (3.1) cannot be computed analytically, we thus have to compute an approximated solution. Figure 3.1 shows the discretization process. We first introduce a triangulation of the domain Ω and then we define a finite element basis over the triangulation. Each finite element basis is a piecewise linear function over the triangulation, which take value one at a node of the triangulation and zero at all the other nodes. We approximate (3.1) in the finite element space, and in particular we obtain an approximate solution of the problem solving a linear system. For an accurate description of the discretization see [3].

In this work we restrict our attention to the special case in which the finite element basis is linear (i.e. each basis is a piecewise linear function) and the triangulation is such that the vertices of the triangles are in correspondence of the data locations $\mathbf{p}_i$. This is a standard setting in many applications.

3.3 Numerical Study of Asymptotic Properties

As shown in [22], the best rate of convergence for general penalized regression estimators over a 2-dimensional domain is

$$\text{MSE} \sim n^{-\frac{p}{p+1}}$$

and is achieved choosing

$$\lambda_n \sim n^{-\frac{p}{2(p+1)}}$$

where p is the number of existing derivatives of the function f_0 that we want to estimate. Since the estimator of SR-PDE is searched in the space $H^2(\Omega)$, in our simulations we set $p = 2$ and let λ_n decrease as $n^{-1/3}$. We thus expect to observe a rate of convergence for the bias of the estimator of order $n^{-1/3}$ and for the MSE of order $n^{-2/3}$.

We consider four different simulation settings that are characterized by different boundary conditions (b.c.): Dirichlet exact b.c., Dirichlet wrong b.c., Neuman exact b.c. and Neuman wrong b.c.. In this way, we can also explore the effect of different boundary conditions on the rate of decay of the error. Exact b.c. corresponds to a complete knowledge of α, that is of the phenomenon at the boundary, while wrong b.c. corresponds to no-knowledge of the behaviour at the boundary. The error is computed in the discrete norm on the data locations. We use the same spatial domain and the same test function considered in the first chapter of [1], where the convergence is studied in the case of exact Dirichlet boundary conditions.

Figures 3.2 and 3.3 show the bias of the SR-PDE estimator with respect to the number of observations n in case of Dirichlet and Neuman boundary conditions respectively. To compute the bias the method is applied to the exact data, without adding any noise at the evaluations. We can observe that both in the Dirichlet and Neuman case the expected rate of convergence is achieved in case of exact boundary conditions. The rate on decay of the bias is strongly influenced by wrong Dirichlet boundary conditions, as we can observe from Figure 3.2 the error is practically non decrising for large values of n. Wrong Neuman boundary conditions still affect the rate of decay of the bias, however, as we can observe from Figure 3.3, the bias is still decreasing for large values of n.

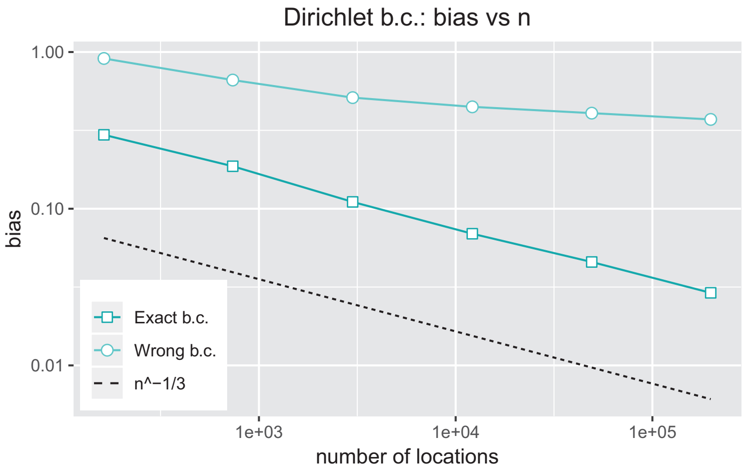


Fig. 3.2 Test functions without noise; exact and wrong Dirichlet boundary conditions. Convergence rates of the bias of the finite element estimator with respect to the number of observations n, with $\lambda_n = n^{-2/3}$.

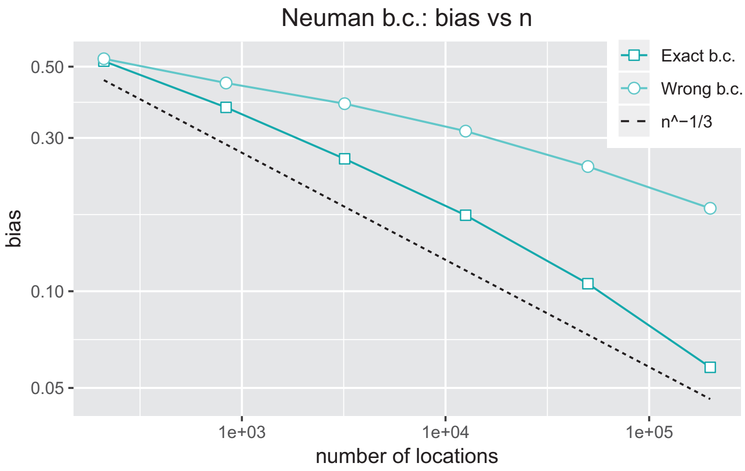


Fig. 3.3 Test functions without noise; exact and wrong Neuman boundary conditions. Convergence rates of the bias of the finite element estimator with respect to the number of observations n, with $\lambda_n = n^{-2/3}$.

Figures 3.4 and 3.5 show the MSE of the SR-PDE estimator with respect to the number of observations n in case of Dirichlet and Neuman boundary conditions respectively. To compute the MSE a gaussian incorrelated noise is added to the exact data. As for the bias, we observe that wrong Dirichlet boundary conditions strongly affect the performance of the estimator. The expected rate is achieved in the exact Dirichlet and in the wrong Neuman case. In the exact Neuman case the rate of convergence seems to be faster than expected, this may be due to the fact that, even if the estimator is searched in the space $H^2(\Omega)$, the true f_0 has more than two derivatives.

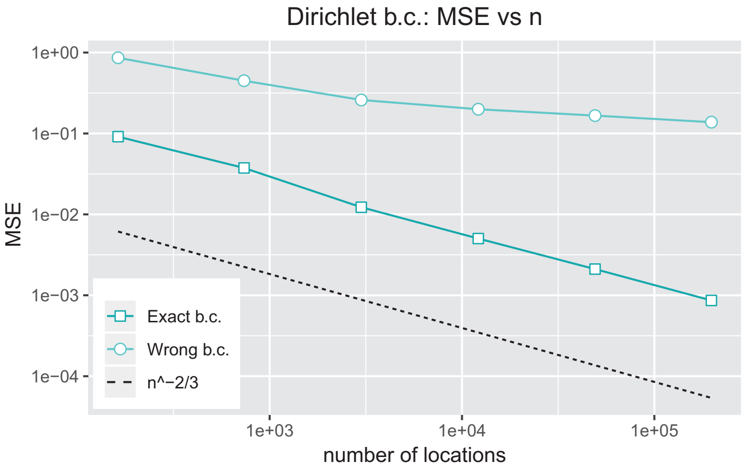


Fig. 3.4 Data with noise; exact and wrong Dirichlet boundary conditions. Convergence rates of the MSE of the finite element estimator with respect to the number of observations n, with $\lambda_n = n^{-2/3}$.

3.4 Future Directions

We have numerically investigated the rate of decay of the bias and the MSE of the SR-PDE estimator, showing that the optimal rate of convergence can be achieved when Dirichlet or Neuman exact boundary conditions are enforced. We also have shown that wrong Neuman boundary conditions affect the rate of decay of the error, that however continue to decay for large values of n. The empirical results displayed in this work support the consistency of SR-PDE estimator. We are currently working on proving the consistency theoretically.

We have here considered a standard choice of the discretization of the domain, with a finite element basis for each data location. However, the SR-PDE does not impose this restriction. An interesting future development is the study of the rate of convergence when the finite element basis is not constrained to the data locations, in

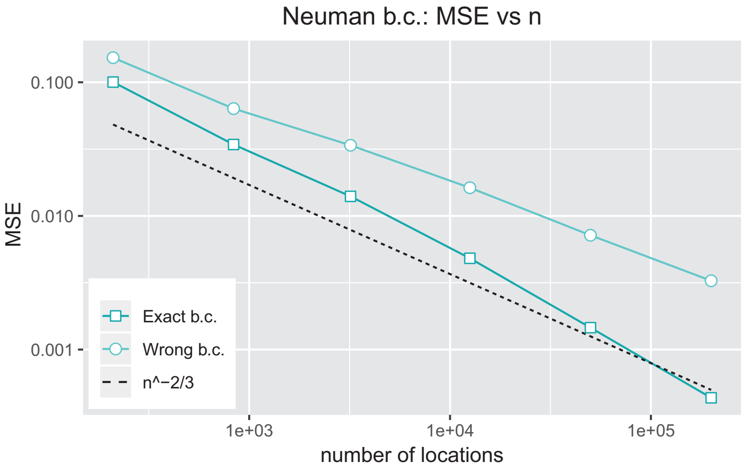


Fig. 3.5 Data with noise; exact and wrong Neuman boundary conditions. Convergence rates of the MSE of the finite element estimator with respect to the number of observations n, with $\lambda_n = n^{-2/3}$.

order to have a finer or coarser triangulation of the domain, that may not directly be linked to the number of observations.

References

[1] Arnone, E.: Regression with PDE penalization for modelling functional data with spatial and spatio-temporal dependence. Doctoral dissertation (2018)

[2] Azzimonti, L., Nobile, F., Sangalli, L.M., Secchi, P.: Mixed Finite Elements for Spatial Regression with PDE Penalization. SIAM/ASA Journal on Uncertainty Quantification **2**(1), 305–335 (2014)

[3] Azzimonti, L., Sangalli, L.M., Secchi, P., Domanin, M., Nobile, F.: Blood flow velocity field estimation via spatial regression with PDE penalization. Journal of the American Statistical Association **110**(511), 1057–1071 (2015)

[4] Cox, D.D.: Asymptotics for M-type smoothing splines. The Annals of Statistics, 530–551 (1983)

[5] Cox, D.D.: Multivariate smoothing spline functions. SIAM Journal on Numerical Analysis **21**(4), 789–813 (1984)

[6] Cucker, F., Zhou, D.X: Learning theory: an approximation theory viewpoint (Vol. 24). Cambridge University Press (2007)

[7] Delicado, P., Giraldo, R., Comas, C., Mateu, J.: Statistics for spatial functional data: some recent contributions. Environmetrics **21**(3-4), 224–239 (2010)

[8] Ettinger, B., Guillas, S., Lai, M.J.: Bivariate splines for ozone concentration forecasting. Environmetrics **23**(4), 317–328 (2012)

[9] Eubank, R.L.: Nonparametric regression and spline smoothing. CRC press (1999)

[10] Ferraty, F., Vieu, P.: Nonparametric functional data analysis: theory and practice. Springer Science & Business Media (2006)

[11] Guillas, S., Lai, M.J.: Bivariate splines for spatial functional regression models. Journal of Nonparametric Statistics **22**(4), 477–497 (2010)

[12] Györfi, L., Kohler, M., Krzyzak, A., Walk, H.: A distribution-free theory of nonparametric regression. Springer Science & Business Media (2006)

[13] Huang, J.Z: Local asymptotics for polynomial spline regression. The Annals of Statistics **31(5)**, 1600–1635 (2003)

[14] Kokoszka, P., Reimherr, M.: Introduction to functional data analysis. Chapman and Hall/CRC (2017)

[15] Lai, M.J., Schumaker, L.L.: Spline functions on triangulations (No. 110). Cambridge University Press (2007)

[16] Lai, M.J., Wang, L.: Bivariate penalized splines for regression. Statistica Sinica **23**(3), 1399–1417 (2013)

[17] Mateu, J., Romano, E.: Advances in spatial functional statistics. Stoch Environ Res Risk Assess **31**, 1–6 (2017)

[18] Ramsay, T.: Spline smoothing over difficult regions. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **54**(2), 307–319 (2002)

[19] Ramsay, J., Silverman, B.: Functional Data Analysis (Second edition). Springer, New York (2005)

[20] Sangalli, L.M. and Ramsay, J.O., Ramsay, T.O.: Spatial spline regression models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **75**(4), 681–703 (2013)

[21] Scott-Hayward, L.A.S., Mackenzie, M.L., Donovan, C.R., Walker, C.G., Ashe, E.: Complex region spatial smoother (CReSS). Journal of Computational and Graphical Statistics **23**(2), 340–360 (2014)

[22] Stone, C.J.: Optimal global rates of convergence for nonparametric regression. The annals of statistics, 1040–1053 (1982)

[23] Wang, H., Ranalli, M.G.: Low-rank smoothing splines on complicated domains. Biometrics **63**(1), 209–217 (2007)

[24] Wood, S.N., Bravington, M.V., Hedley, S.L.: Soap film smoothing. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70**(5), 931–955 (2008)

Chapter 4
Learning with Signatures

Gérard Biau and Adeline Fermanian

Abstract Sequential and temporal data arise in many fields of research, such as quantitative finance, medicine, or computer vision. The present article is concerned with a novel approach for sequential learning, called the signature method and rooted in rough path theory. Its basic principle is to represent multidimensional paths, i.e., functions from $[0, 1]$ to $\mathbb{R}^d$, by a graded feature set of their iterated integrals, called the signature. This approach relies critically on an embedding principle, which consists in representing discretely sampled data as continuous paths. After a survey of basic principles of signatures, we investigate the influence of embeddings on prediction accuracy with an in-depth study of recent and challenging datasets. We show that a specific embedding, called lead-lag, is systematically better, whatever the dataset or algorithm used.

4.1 Introduction

Sequential or temporal data occur in many fields of research, due to an increase in storage capacity and to the rise of machine learning techniques. Sequential data are characterized by the fact that each sample consists of an ordered array of values. Although the ordering often corresponds to time, it is not always the case. For example, text documents or DNA sequences have an intrinsic ordering, and can, therefore, be considered as sequential. Besides, when time is involved, several values can be recorded simultaneously, giving rise to an ordered array of vectors, which is, in the field of time series, often referred to as multidimensional time series. To name only a few domains, market evolution is described by financial time series,

Gérard Biau (✉)
Sorbonne Université, Laboratoire de Probabilités, Statistique et Modélisation, 4 place Jussieu, 75005 Paris, France, e-mail: gerard.biau@sorbonne-universite.fr

Adeline Fermanian
Sorbonne Université, Laboratoire de Probabilités, Statistique et Modélisation, 4 place Jussieu, 75005 Paris, France, e-mail: adeline.fermanian@sorbonne-universite.fr

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_4

and physiological variables (e.g., electrocardiograms, electroencephalograms...) are recorded simultaneously in medicine, yielding multidimensional time series. We can also mention smartphone and GPS sensors data, or character recognition problems, where data has both a spatial and temporal aspect. These high-dimensional datasets open up new theoretical and practical challenges, as both statistical models and algorithms need to be adapted to their sequential nature.

Our goal in the present article is to discuss a novel approach for sequential learning, called the signature method, and coming from rough path theory. Its main idea is to summarize temporal (or functional) inputs by the graded feature set of their iterated integrals, the signature. Note that, in rough path theory, functions are referred to as paths, to insist on their geometrical aspects. Indeed, the importance of iterated integrals had been noticed by geometers in the 60s, as presented in the seminal work of [2]. It has been rediscovered by [10] in the context of stochastic analysis and controlled differential equations, and is at the heart of rough path theory. This theory, of which [11] and [5] give a recent account, focuses on developing a new notion of paths to make sense of evolving irregular systems. In this context, it has been shown that the signature provides an accurate summary of a path and allows to obtain arbitrarily good linear approximations of continuous functions of paths. Therefore, assuming we want to learn an output $Y \in \mathbb{R}$, which depends on a random path $X : [0,1] \to \mathbb{R}^d$, rough path theory suggests that the signature is a relevant feature set to describe X.

As can be expected, the signature has recently received the attention of the machine learning community and has achieved a series of successful applications. To cite some of them, [14] have achieved state-of-the-art results for handwriting recognition with a recurrent neural network combined with signature features. [7] have used the same approach for character recognition, and [8] have coupled Lasso with signature features for financial data streams classification. [1] have investigated its use for the detection of bipolar disorders, and [15] for human action recognition. For a gentle introduction to the signature method in machine learning, we refer the reader to [3].

However, despite many promising empirical successes, a lot of questions remain open, both practical and theoretical. In particular, to compute the signature, it is necessary to embed discretely sampled data points into paths. While authors use different approaches, this embedding is only mentioned in some articles, and rarely discussed. Thus, our purpose in this paper is to take a step forward in understanding how signature features should be constructed for machine learning tasks, with a special focus on the embedding step.

Our document is organized as follows. First, in Section 4.2, we give a brief exposition of the signature definition and properties. Then, we compare the predictive performance of different embeddings in Section 4.3. We emphasize that the embedding is as a crucial step as the algorithm choice since it can drastically change accuracy results. Moreover, we point out that one embedding, called lead-lag, performs systematically better than others, and this consistently over different datasets and learning algorithms.

4.2 Signature Definition and First Properties

We introduce in this section the notion of signature and review some of its important properties. The reader is referred to [11] or [5] for a more involved mathematical treatment with proofs. Throughout the article, our basic objects are paths, that is functions from $[0, 1]$ to $\mathbb{R}^d$, where $d \in \mathbb{N}^*$. The main assumption is that these paths are of bounded variation, i.e., they have finite length.

Definition 1 Let

$$\begin{aligned} X : [0,1] &\longrightarrow \mathbb{R}^d \\ t &\longmapsto (X_t^1, \dots, X_t^d). \end{aligned}$$

The total variation of X is defined by

$$\|X\|_{1-var} = \sup_D \sum_{t_i \in D} \|X_{t_i} - X_{t_{i-1}}\|,$$

where the supremum is taken over all finite partitions

$$D = \big\{(t_0, \dots, t_k) \mid k \geq 1,\ 0 = t_0 < t_1 < \cdots < t_{k-1} < t_k = 1\big\}$$

of $[0, 1]$, and $\| \cdot \|$ denotes the Euclidean norm on $\mathbb{R}^d$. The path X is said to be of bounded variation if its total variation is finite.

The assumption of bounded variation allows to define Riemann-Stieljes integrals along paths. From now on, we assume that the integral of a continuous path $Y : [0, 1] \to \mathbb{R}^d$ against a path of bounded variation $X : [0, 1] \to \mathbb{R}^d$ is well-defined on any $[s, t] \subset [0, 1]$, and denoted by

$$\int_s^t Y_u dX_u = \begin{pmatrix} \int_s^t Y_u^1 dX_u^1 \\ \vdots \\ \int_s^t Y_u^d dX_u^d \end{pmatrix} \in \mathbb{R}^d,$$

where $X = (X^1, \dots, X^d)$ and $Y = (Y^1, \dots, Y^d)$. We are now in a position to define the signature.

Definition 2 Let $X : [0, 1] \to \mathbb{R}^d$ be a path of bounded variation, $I = (i_1, \dots, i_k) \subset \{1, \dots, d\}^k$, $k \in \mathbb{N}^*$, be a multi-index of length k, and $[s, t] \subset [0, 1]$ be an interval. The signature coefficient of X corresponding to the multi-index I on $[s, t]$ is defined by

$$S^I(X)_{[s,t]} = \int \cdots \int_{s \leq u_1 < \cdots < u_k \leq t} dX_{u_1}^{i_1} \dots dX_{u_k}^{i_k}. \tag{4.1}$$

$S^I(X)_{[s,t]}$ is then said to be a signature coefficient of order k.

The signature of X is the sequence containing all signature coefficients, i.e.,

$$S(X)_{[s,t]} = \big(1, S^{(1)}(X)_{[s,t]}, \dots, S^{(d)}(X)_{[s,t]}, S^{(1,1)}(X)_{[s,t]}, S^{(1,2)}(X)_{[s,t]}, \dots\big).$$

The signature of X truncated at order K, denoted by $S^K(X)$, is the sequence containing all signature coefficients of order lower than or equal to K, that is

$$S^K(X)_{[s,t]} = \big(1, S^{(1)}(X)_{[s,t]}, S^{(2)}(X)_{[s,t]}, \dots, S^{\overbrace{(d, \dots, d)}^{K}}(X)_{[s,t]}\big).$$

For simplicity, when $[s,t] = [0,1]$, we omit the interval in the notations, and, e.g., write $S^K(X)$ instead of $S^K(X)_{[0,1]}$. We note that, for a path in $\mathbb{R}^d$, there are d^k coefficients of order k. The signature truncated at order K is therefore a vector of dimension

$$\sum_{k=0}^{K} d^k = \frac{d^{K+1} - 1}{d - 1} \quad \text{if } d \neq 1,$$

and $K + 1$ if $d = 1$. Unless otherwise stated, we assume that $d \neq 1$, as this is in practice usually the case. Thus, the size of $S^K(X)$ increases exponentially with K, and polynomially with d. Finally, it should be noted that, due to the ordering in the integration domain in (4.1), signature coefficients are not symmetric. For example, $S^{(1,2)}(X)$ is a priori not equal to $S^{(2,1)}(X)$.

A crucial feature of the signature is that it encodes geometric properties of the path. Indeed, it is clear that coefficients of order 2 correspond to some areas outlined by the path, as shown in Figure 4.1. For higher orders of truncation, the signature contains information about the joint evolution of tuples of coordinates. Furthermore, the signature possesses several properties that make it a good statistical summary of paths, as shown in the next three propositions.

Proposition 1 *Let $X : [0,1] \to \mathbb{R}^d$ be a path of bounded variation, and $\psi : [0,1] \to [0,1]$ be a non-decreasing surjection. Then, if $\widetilde{X}_t = X_{\psi(t)}$ is the reparametrization of X under ψ,*

$$S(\widetilde{X}) = S(X).$$

In other words, the signature of a path is the same up to any reasonable time change. There is, therefore, no information about the path travel time in signature coefficients, which may be a useful feature in some applications. Nevertheless, when relevant for the problem at hand, it is possible to include this information by adding the time parametrization as a coordinate of the path. A second important property is a condition ensuring uniqueness of signatures.

Proposition 2 *If X has at least one monotonous coordinate, then $S(X)$ determines X uniquely.*

It should be noticed that having a monotonous coordinate is a sufficient condition, but a necessary one can be found in the monograph by [9], together with a proof of the proposition. The principal significance of this result is that it provides a practical procedure to guarantee signature uniqueness: it is sufficient to add a monotonous coordinate to the path X. For example, the time embedding mentioned above will

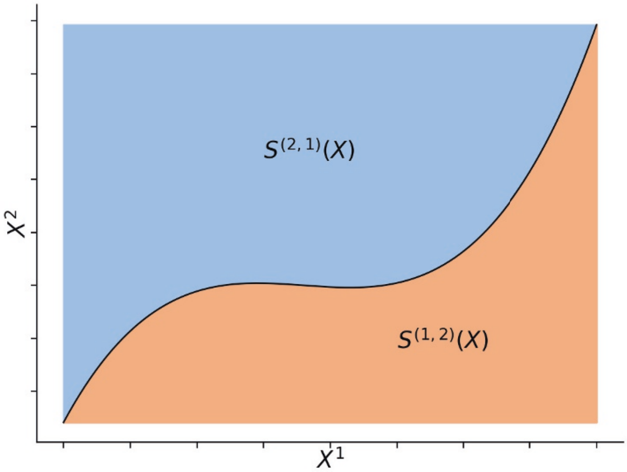


Fig. 4.1 Geometric interpretation of signature coefficients.

satisfy this condition. The next proposition reveals that the signature linearizes functions of X.

Proposition 3 *Let D be a compact subset of the space of bounded variation paths from $[0,1]$ to $\mathbb{R}^d$ that are not tree-like equivalent. Let $f : D \to \mathbb{R}$ be continuous. Then, for every $\epsilon > 0$, there exists $N \in \mathbb{N}^*$, $w \in \mathbb{R}^N$, such that, for any $X \in D$,*

$$\big| f(X) - \langle w, S(X) \rangle \big| \leq \epsilon,$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean scalar product on $\mathbb{R}^N$.

The notion of tree-like equivalence is closely related to the uniqueness of paths—the reader is referred to [9] for a definition. Proposition 3 is then a consequence of the Stone-Weierstrass theorem.

4.3 Embeddings

Now that we have presented the signature and its properties, we focus on its use in machine learning. In this context, we place ourselves in a statistical framework, and assume that our goal is to understand the relationship between a random input path $X : [0,1] \to \mathbb{R}^d$ and a random output $Y \in \mathbb{R}$. In a classical setting, we would be given a sample of independent and identically distributed (i.i.d.) observations $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, drawn from (X, Y). However, in applications, we only observe a realization X_i sampled at a discrete set of times $0 \leq t_1 < \cdots < t_{p_i} \leq 1$,

$p_i \in \mathbb{N}^*$. Therefore, we are given an i.i.d. sample $\{(\mathbf{x_1}, Y_1), \ldots, (\mathbf{x_n}, Y_n)\}$, where $\mathbf{x_i}$ takes the form of a matrix, i.e.,

$$\mathbf{x_i} = \begin{pmatrix} x^1_{i,1} & \cdots & x^1_{i,p_i} \\ \vdots & & \vdots \\ x^d_{i,1} & \cdots & x^d_{i,p_i} \end{pmatrix} \in \mathbb{R}^{d \times p_i}. \tag{4.2}$$

In this notation, $x^k_{i,j}$ denotes the kth coordinate of the ith sample observed at time t_j. If $d = 1$, we are in a classical setting of time series, where each observation is sampled in a finite number of points. However, d may here differ from 1, so we find ourselves in a more general situation where we want to learn from multidimensional time series. Moreover, it is worth noting the dependence of the number of sampled points p_i on i. In other words, each observation may have a different length. The signature dimension being independent of the number of sampled points, representing time series by their signature naturally handles inputs of various lengths, whereas traditional methods often require them to be normalized to a fixed length. To sum up, the signature method is appropriate for learning with discretely sampled multidimensional time series, possibly of different lengths.

To use signature features, one needs to embed the observations $\mathbf{x_i}$ into paths of bounded variation $X_i : [0, 1] \to \mathbb{R}^d$. Therefore, we need to choose an interpolation method, but, to ensure some properties such as signature uniqueness (see Proposition 2), we may also create new coordinates to the path and in this way increase the dimension of the embedding space. We refer the reader to [4] for a detailed description of the different embeddings that we use in the present article.

Our empirical study is based on three datasets of various nature. We present here the results on the Quick, Draw! dataset but similar results are obtained on two other datasets, described in [4]. The Quick, Draw! dataset has been made available by Google [6], and consists of drawing trajectories. It is made up of 50 million drawings, each drawing being a sequence of time-stamped pen stroke trajectories, divided into 340 categories. Some samples are shown in Figure 4.2.

We present in Figure 4.3 the results of our study on embedding performance, obtained with the following approach. Starting from the raw data, we first embed it into a continuous path, then compute its truncated signature, and use this vector as input for a learning algorithm. We want our findings to be independent of the data and the underlying statistical model so we use a range of different algorithms. The classification metric to assess prediction quality is the accuracy score. Then, to compare the quality of different embeddings, we plot the accuracy score against the log number of features, which yields one curve per embedding, where each point corresponds to a different truncation order. We then check whether one embedding curve is above the others, which would mean that, at equal input size, this embedding is better for learning.

A first striking fact is that some embeddings, namely the time and lead-lag, seem consistently better, whatever the algorithm used. It suggests that this performance is due to intrinsic theoretical properties of signatures and embeddings, not to domain-specific characteristics. The linear and rectilinear embeddings (red and pink curves),

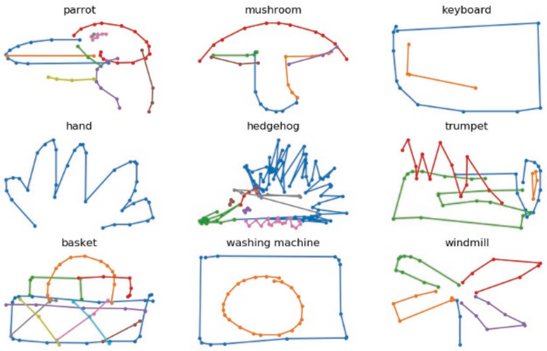


Fig. 4.2 9 samples from the Quick, Draw! dataset. Each color corresponds to a different pen stroke.

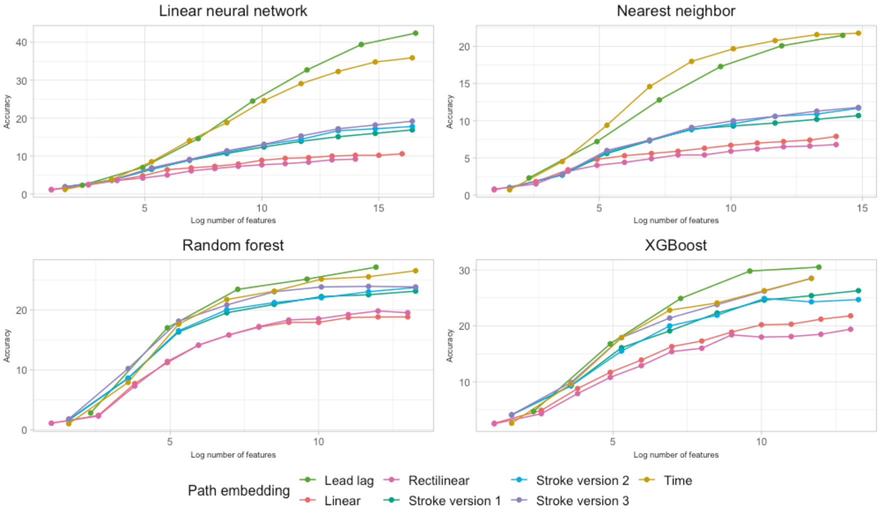


Fig. 4.3 Quick, Draw! dataset: prediction accuracy on the test set, for different algorithms and embeddings.

which are often used in the literature, appear to give the worst results. This bad performance can be explained by the fact that there is no guarantee that the signature characterizes paths when using the linear or rectilinear embeddings. Therefore, two different paths can have the same signature, without necessarily corresponding to the same class.

To conclude, the take-home message is that using the lead-lag embedding seems to be the best choice, regardless of the data and algorithm used. It does not cost anything computationally and can drastically improve prediction accuracy. Moreover, the linear and stroke paths yield surprisingly poor results, despite their frequent use in the literature.

References

[1] Arribas, I.P., Goodwin, G.M., Geddes, J.R., Lyons, T., Saunders, K.E.: A signature-based machine learning model for distinguishing bipolar disorder and borderline personality disorder. Translational psychiatry **8**, 274 (2018)

[2] Chen, K.: Integration of paths—a faithful representation of paths by non-commutative formal power series. Transactions of the American Mathematical Society **89**, 395–407 (1958)

[3] Chevyrev, I., Kormilitzin, A.: A primer on the signature method in machine learning. arXiv:1603.03788 (2016)

[4] Fermanian, A.: Embedding and learning with signatures. arXiv:1911.13211 (2019)

[5] Friz, P., Victoir, N.: Multidimensional Stochastic Processes as Rough Paths: Theory and Applications. Cambridge University Press, Cambridge (2010)

[6] Google Creative Lab: The Quick, Draw! Dataset. (2017) Available from https://github.com/googlecreativelab/quickdraw-dataset

[7] Graham, B.: Sparse arrays of signatures for online character recognition. arXiv:1308.0371 (2013)

[8] Gyurkó, L., Lyons, T., Kontkowski, M., Field, J.: Extracting information from the signature of a financial data stream. arXiv:1307.7244 (2014)

[9] Hambly, B., Lyons, T.: Uniqueness for the signature of a path of bounded variation and the reduced path group. Annals of Mathematics **171**, 109–167 (2010)

[10] Lyons, T.: Differential equations driven by rough signals. Revista Matemática Iberoamericana **14**, 215–310 (1998)

[11] Lyons, T., Caruana, M., Lévy, T.: Differential Equations Driven by Rough Paths. Springer, Berlin (2007)

[12] Malekzadeh, M., Clegg, R.G., Cavallaro, A., Haddadi, H.: Protecting sensory data against sensitive inferences. In: Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems, ACM (2018)

[13] Salamon, J., Jacoby, C., Bello, J.P.: A dataset and taxonomy for urban sound research. In: Proceedings of the 22nd ACM International Conference on Multimedia, 1041–1044. ACM (2014)

[14] Yang, W., Jin, L., Liu, M.: Deepwriterid: An end-to-end online text-independent writer identification system. IEEE Intelligent Systems **31**, 45–53 (2016)

[15] Yang, W., Lyons, T., Ni, H., Schmid, C., Jin, Li., Chang, J.: Leveraging the path signature for skeleton-based human action recognition. arXiv:1707.03993 (2017)

Chapter 5
About the Complexity Function in Small-ball Probability Factorization

Enea G. Bongiorno, Aldo Goia and Philippe Vieu

Abstract The Small-Ball Probability (SmBP) of a process valued in a semi-metric space is considered. Assume that it factorizes in two terms that play the role of a surrogate density and of a volumetric term, respectively. This work presents some recent developments concerning the study of the volumetric term that detains information about the complexity of the underlying process. In particular, once some estimators and their asymptotics are presented, a goodness-of-fit multiple testing procedure is implemented in order to detect the complexity family the process belongs to.

5.1 Introduction

Functional statistics has received a lot of attention in the recent years reaching a good maturity level that has produced a series of interesting monographs; see, as an instance [10, 13, 15, 19]. An interesting issue in functional statistics is to evaluate the complexity extent of the probability law of a random process given a sample of discretized trajectories. So far, different approaches can be found in literature; they share the idea of measuring some (fractal) dimension of the process such as correlation or Hausdorff dimension: as an instance, see [1, 8, 14], and more recently [4, 5, 6, 7].

In some way, all the introduced methodologies are based on the concept of small ball probability: given a random element X valued in a suitable semimetric space $\mathcal{F}$ and denoting by $B(\chi, h)$ the ball centered at $\chi \in \mathcal{F}$ with radius $h > 0$, the small

Enea G. Bongiorno (✉)
Università del Piemonte Orientale, Via Perrone 18, 28100, Novara, Italy,
e-mail: enea.bongiorno@uniupo.it

Aldo Goia
Università del Piemonte Orientale, Via Perrone 18, 28100, Novara, Italy,
e-mail: aldo.goia@uniupo.it

Philippe Vieu
Institut de Mathématiques de Toulouse, Université Paul Sabatier, Toulouse, France,
e-mail: philippe.vieu@math.univ-toulouse.fr

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_5

ball probability of X is $\mathbb{P}\left(X \in B\left(\chi, h\right)\right)$ when h tends to zero. In many situations it is convenient to suppose that

$$\mathbb{P}\left(X \in B\left(\chi, h\right)\right) \sim \psi\left(\chi\right) \phi\left(h\right) \qquad \text{as } h \to 0, \tag{5.1}$$

where, to ensure the identifiability of the decomposition, a normalising restriction is necessary, such as for instance

$$\mathbb{E}\left[\psi(X)\right] = 1. \tag{5.2}$$

Factorization (5.1) is not a forcing assumption; indeed it holds true under appropriate hypotheses (see, for instance, [3, 18]). The convenience in assuming (5.1) is, at least, twofold. Firstly, the function $\psi\left(\chi\right)$ can be interpreted as a surrogate density of the functional random element X and exploited in different frameworks; the interested reader can appreciate its potential by looking, as an instance, at [2, 9, 11] where the surrogate density is estimated in different ways and employed to define a notion of mode or for classification purposes. Secondly, the function $\phi\left(h\right)$ plays the role of the volumetric term and can be used to evaluate the complexity of the probability law of the process X (see [5]). To fix the ideas, note that for some special families of processes it is possible to specify the complexity function $\phi(h)$ in a parametric form by means of some complexity index $\theta \in \mathbb{R}^p$ (p being a positive integer). For example, if the process has a fractal structure (see [10, Definition 13.1]) then $\phi_\theta(h) = c_\theta h^\theta$, for a constant term c_θ and $\theta > 0$. Another notable example comes from infinite dimensional Gaussian processes (see [18]), for which $\phi_\theta(h) = C_1 h^\alpha \exp\left\{-C_2/h^\beta\right\}$ with $\theta = (\alpha, \beta) \in [0, \infty) \times (0, \infty)$ and positive constants C_1, C_2. In other words, the form of the volumetric term provides information related to the complexity of the probability law of the process and the parameter θ can be interpreted as a measure of its complexity.

This paper summarizes some recent efforts [4, 5, 6, 9] in studying the complexity factor ϕ and complexity parameter θ and illustrates some new results (see [7]). In particular, Section 5.2 furnishes a couple of nonparametric estimators of ϕ and illustrates their asymptotic properties: (uniform) consistency and asymptotic normality. In Section 5.3, it is assumed that ϕ is parametrically specified, and then an estimator of the complexity parameter θ is presented together with its asymptotic properties (weak consistency and normality). Finally, in Section 5.4 a goodness-of-fit test, based on a multiple testing procedure, for the complexity term ϕ is illustrated.

5.2 Nonparametric Estimators of the Complexity Factor

In the literature the estimation problem of the complexity factor $\phi(h)$ has been already treated. More in detail, given a sample of n discretized curves $X_1, \dots, X_n$ drawn from X, and $h > 0$ sufficiently close to zero, the simplest estimator is the empirical one

$$\widehat{\phi}_{\text{emp}}\left(h\right) = \frac{1}{n\left(n-1\right)} \sum_{j=1}^{n} \sum_{i \neq j} 1_{B\left(X_j, h\right)}\left(X_i\right) \tag{5.3}$$

proposed in [6], where $1_A(x)$ is the characteristic function of the set A.

Previously, a kernel based estimator was proposed in [9] as a by-product of the estimation of the surrogate-density ψ. It was slightly modified (averaging over the sample) in [5] to avoid the dependence on the point at which the SmBP factorization is estimated as it should be according to assumption (5.1). Its final form is as follows:

$$\widehat{\phi}_{\text{kNN}}(h) = \sum_{j=1}^{n} \left\{ \frac{\sum_{i=1}^{n} k_{i,j}}{kn(n-1)} \sum_{i \neq j} 1_{B(X_j,h)}(X_i) \right\}, \tag{5.4}$$

where $k < n$ is a positive integer, $k_{i,j} = \#\left\{l \neq i : X_l \in B\left(X_i, H_{n,k}(X_j)\right)\right\}$ and $H_{n,k}(X_j) = \min\left\{h \in \mathbb{R}^+, \sum_{i=1}^{n} 1_{B(X_j,h)}(X_i) = k\right\}$.

Before looking at the theoretical properties of these two estimators, it is worth to notice that from a practical point of view, because the asymptotic factorization (5.1) holds for small h, too large values must be avoided since they may increase the estimation error. At the same time, also too small values of h must be discarded since, for small sample sizes, they force $\widehat{\phi}$ to be null: indeed the ball $B\left(X_j, h\right)$ could contain no sample points other than X_j. In other words, in practice a suitable range of values $\mathcal{H} = [h_m, h_M]$ for h should be identified (have a look at [5, 6] for some data driven ideas on this issue).

5.2.1 Some Asymptotics

From a theoretical point of view, the above estimators have good asymptotic properties. In particular, assume that

(A.1) for any $h > 0$, $\mathbb{P}(X \in B(\chi, h)) > 0$;
(A.2) the model defined by (5.1) and (5.2) holds;
(A.3) ϕ is increasing on a neighbourhood of zero, strictly positive and tends to zero as h goes to zero;
(A.4) ψ is bounded and $\psi(\chi) > 0$.

Thus, estimators (5.3) and (5.4) satisfies the following proposition.

Proposition 1 *Under (A.1)–(A.4),*

- *(see [9, Corollary 5.1]) the estimator $\widehat{\phi}_{\text{kNN}}(h)$ converges in probability to $\phi(h)$ as $n \longrightarrow \infty$.*
- *(see [6, Proposition 1]) the estimator $\widehat{\phi}_{\text{emp}}(h)$ is asymptotically unbiased with variance*

$$\text{Var}\left(\widehat{\phi}_{\text{emp}}(h)\right) = \frac{4(n-2)}{n(n-1)}\sigma_1^2(h) + \frac{2}{n(n-2)}\sigma_2^2(h) \tag{5.5}$$

where $\sigma_1^2(h) = \text{Var}\left(\mathbb{E}\left[1_{\{X_1 \in B(X_2,h)\}}\middle| X_2\right]\right)$ and $\sigma_2^2(h) = \text{Var}\left(1_{\{X_1 \in B(X_2,h)\}}\right)$ are positive and finite. Moreover, its standardized version converges in law to a standard Gaussian distribution as $n \longrightarrow \infty$.

The proof of the latter proposition is strictly related to the fact that both $\widehat{\phi}_{\mathrm{kNN}}(h)$ and $\widehat{\phi}_{\mathrm{emp}}(h)$ can be seen as two-order U-statistics and classical asymptotic results can be applied (see [16, 17]). More theoretical details are given in [9, 6].

The next proposition provides the uniform (in $\mathcal{H}$) consistency of a large class of nonparametric estimators of ϕ that, as particular cases, includes also (5.3) and (5.4).

Proposition 2 *(See [7, Proposition 1]) Assume (A.1)–(A.4). Let $\widehat{\phi}$ be an estimator of ϕ such that $\widehat{\phi}(h) \longrightarrow \phi(h)$ for any $h \in \mathcal{H}$ and such that $\widehat{\phi}(\cdot)$ is an increasing function on $\mathcal{H}$. If ϕ is continuous and increasing on $\mathcal{H}$, then $\widehat{\phi}$ is convergent to ϕ in probability, uniformly on $\mathcal{H}$.*

The proof of the latter is based on the fact that ϕ is uniform continuous on $\mathcal{H}$ and that $\widehat{\phi}$ is pointwise consistent for ϕ over $\mathcal{H}$; the interested reader can find the details in [7].

5.3 Parametric Estimation of the Complexity Factor

Assume that ϕ is specified by some parametric relation of the form

$$\phi \in \{\phi_\theta, \theta \in \Theta \subset \mathbb{R}^p\}. \tag{5.6}$$

Some examples of processes for which a parametric form of ϕ is available, are the fractal and the infinite dimensional Gaussian processes (see the Introduction). In these cases, estimating θ can provide the complexity degree of the underlying process. An estimate θ_n of θ is defined as the minimizer, over a suitable compact subset Θ of $\mathbb{R}^p$, of a dissimilarity measure between the target ϕ_θ and a nonparametric estimator $\widehat{\phi}$ (see Section 5.2). In particular, in [7] the authors consider the *centered cosine dissimilarity* between $g(\phi_\theta)$ and $g(\widehat{\phi})$ computed on the observed values and defined by

$$\Delta(\widehat{\phi}, \phi_\theta) = 1 - \frac{\langle g(\phi_\theta), g(\widehat{\phi})\rangle^2}{\|g(\phi_\theta)\|^2 \|g(\widehat{\phi})\|^2}, \tag{5.7}$$

where $g : (0, \infty) \longrightarrow \mathbb{R}$ is a suitable continuous function that pointwisely transforms $\phi(h)$ in $g(\phi(h))$ for all $h \in \mathcal{H}$, while $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ denote the usual inner product and the associated norm, respectively, of the Hilbert space $\mathcal{L}^2_{\mathcal{H}}$ of square integrable real functions defined on $\mathcal{H}$. If ϕ_θ and $\widehat{\phi}$ are both bounded away from zero on $\mathcal{H}$, then $\phi_\theta(\cdot)$, $\widehat{\phi}(\cdot)$, $g(\phi_\theta(\cdot))$ and $g(\widehat{\phi}(\cdot))$ are in $\mathcal{L}^2_{\mathcal{H}}$ and (5.7) is well-posed. In practice, possible working choices of g are the identity function in the fractal case, and the logarithm for many Gaussian processes (see [5]). Hence, an estimator θ_n of θ is:

$$\theta_n = \arg\min_{\theta \in \Theta} \Delta(\widehat{\phi}, \phi_\theta). \tag{5.8}$$

5.3.1 Some Asymptotics

Some new asymptotic properties of (5.8) have been derived in the [7] where it has been proved that θ_n is a weakly $\sqrt{n}$–consistent estimator of θ and asymptotically

Gaussian distributed.
In particular, consider the following assumptions

(B.1) the model defined by (5.1), (5.2) and (5.6) holds;
(B.2) g is Hölder continuous (i.e. there exist $C < \infty, \beta > 0$, such that for all $y_1, y_2 \in \mathbb{R}, |g(y_1) - g(y_2)| \leq C|y_1 - y_2|^{\beta}$);
(B.3) for each $\theta \in \Theta$, the function $\phi_\theta(\cdot)$ is continuous and increasing on $\mathcal{H}$;
(B.4) $\widehat{\phi}$ is a nonparametric estimator of ϕ being pointwise (with respect to $h \in \mathcal{H}$) consistent and increasing in $\mathcal{H}$ ($\widehat{\phi}_{\text{kNN}}$ and $\widehat{\phi}_{\text{emp}}$ satisfy these properties).

The following consistency result holds true for the estimator θ_n.

Proposition 3 *(See [7, Theorem 1]) Under assumptions (B.1)–(B.4), the estimator θ_n converges in probability to the true value θ_0 of the parameter θ, as the sample size n goes to infinity.*

An idea of the proof of the latter proposition, whose details are in [7], is as follows. Firstly, one has to show that $\delta_n(\theta) = \Delta(\widehat{\phi}, \phi_\theta)$ converges uniformly on Θ towards $\delta(\theta) = \Delta\left(\phi_{\theta_0}, \phi_\theta\right)$ which is, as a function of θ, uniformly continuous on the compact set $\Theta \subset \mathbb{R}^p$. Secondly, one has to prove that $\delta(\theta_n) \to \delta(\theta_0)$ to guarantee that θ_n converges in probability to θ_0.

Once the consistency of θ_n is obtained, it is interesting to study its asymptotic distribution which is given in the next proposition at the cost of some additional regularity conditions for g and $\delta(\theta)$.

Proposition 4 *(See [7, Theorem 2]) Assume (B.1)–(B.4) and consider θ_n as in* (5.8) *where $\widehat{\phi}$ is the empirical estimator* (5.3)*. Suppose that g is $C^2(0, +\infty)$ (i.e. twice derivable with continuous derivatives over $(0, +\infty)$) with non-null first derivative, δ is $C^2(\Theta)$ and strictly convex over Θ, then as $n \to +\infty$*

$$\sqrt{n}\,(\theta_n - \theta_0) \sim \mathcal{N}(0, \Gamma)$$

where Γ is a suitable covariance matrix depending on ϕ_{θ_0}, g and their derivatives.

The proof of the latter proposition is mainly based on a Taylor expansion of $\delta(\theta)$ (that explains the required additional regularity conditions) and on the properties of $\widehat{\phi}$ seen as a U-statistic; the interested reader can found more details in [7].

5.4 Testing the Complexity of a Process

The results illustrated above lead to a goodness-of-fit test to compare the complexity function ϕ of observed functional data with a target model ϕ_0. Such a test was firstly introduced in [6] to which the interested reader can refer for further details.
In particular, consider the following hypothesis

$$\begin{aligned} &H_0 : \phi(h) = \phi_0(h) \qquad \text{for all } h \in \mathcal{H} \\ &H_1 : \exists h : \phi(h) \neq \phi_0(h). \end{aligned}$$

Operatively, the authors have implemented a multiple test by considering m values $h_1, \dots, h_m \in \mathcal{H}$, m marginal null hypotheses $H_0^k : \phi(h_k) = \phi_0(h_k)$, $k = 1, \dots, m$ and test statistics

$$D_k^2 = \frac{(\widehat{\phi}_{\text{emp}}(h_k) - \phi_0(h_k))^2}{\text{Var}(\widehat{\phi}_{\text{emp}}(h_k))}, \qquad k = 1, \dots, m, \tag{5.9}$$

which are asymptotically distributed as a chi-square with one degree of freedom (see [6, Proposition 2]). Thus, the asymptotic p-value associated to H_0^k is $p_k = 1 - C_1^2(d_k^2)$ where C_1^2 is the pdf of the r.v. $\chi^2(1)$ and d_k^2 is an observed value of the test statistic. Finally, the decision rule of the multiple test is based on the Holm-Bonferroni correction (see [12]): order p-values $p_{(1)} \leq \cdots \leq p_{(m)}$ and reject H_0 if $p_{(k)} \leq \alpha/(m + 1 - k)$ for at least one k.

In order to operationalize the above test some issues should be addressed in advance. Firstly, a direct calculation for the variance $\text{Var}(\widehat{\phi}_{\text{emp}}(h_k))$, appearing in the denominator of (5.9), is hard to obtain and a resampling technique can be implemented to get an estimate (in [6] a Jackknife estimator is computed). Secondly, in general, the exact expression for $\phi_0(h)$ is rarely available and it can be estimated from an artificial sample generated according to the benchmark model which is supposed to be true. Consequently, the new test statistics become

$$\widetilde{D}_k^2 = \frac{\left(\widehat{\phi}_{\text{emp}}(h_k) - \widehat{\phi}_0(h_k)\right)^2}{V_{n,k} + V_{0,k}}, \qquad k = 1, \dots, m$$

with $V_{n,k}$ and $V_{0,k}$ being some resampling estimators (e.g. the Jackknife ones) of the variances of $\widehat{\phi}_{\text{emp}}$ and $\widehat{\phi}_0$ evaluated at h_k. With these choices, the test statistics $\widetilde{D}_k^2$ are still asymptotically distributed as a chi-square with one degree of freedom and then the multiple test can be implemented as described above.

Finally, as shown in [6], such a test has provided good performances on finite sample size experiments, under various scenarios, through Monte Carlo simulations.

Acknowledgements Authors want to thank the anonymous referee and the editors whose careful reading and suggestions have improved the presentation of the results.
E.G. Bongiorno and A. Goia thank Università del Piemonte Orientale for the financial support and Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

References

[1] Bardet, J.-M.: Tests d'autosimilarité des processus gaussiens. Dimension fractale et dimension de corrélation, Thèse, Université Paris-Sud, France (1997)

[2] Bongiorno, E.G., Goia, A.: Classification methods for Hilbert data based on surrogate density. Comput. Statist. Data Anal. **99**, 204–222 (2016)

[3] Bongiorno, E.G., Goia, A.: Some insights about the small ball probability factorization for Hilbert random elements. Statist. Sinica **27**(4), 1949–1965 (2017)

[4] Bongiorno, E.G., Goia, A., Vieu, P.: On the geometric Brownian motion assumption for financial time series. In: Aneiros G. et. al. (eds) Functional statistics and related fields, 59–65, Contrib. Stat. Springer, Cham (2017)

[5] Bongiorno, E.G., Goia, A., Vieu, P.: Evaluating the complexity of some families of functional data. SORT **42**(1), 27–44 (2018)

[6] Bongiorno, E.G., Goia, A., Vieu, P.: Modeling functional data: a test procedure. Comput. Statist. **34**(2), 451–468 (2019)

[7] Bongiorno, E.G., Goia, A., Vieu, P.: Estimating the complexity index of functional data: some asymptotics. Statist. Probab. Lett. **161**, (2020)

[8] Cutler, C.D., Dawson, D.A.: Nearest-neighbor analysis of a family of fractal distributions. Ann. Probab. **18**, 256–271 (1990)

[9] Ferraty, F., Kudraszow, N., Vieu, P.: Nonparametric estimation of a surrogate density function in infinite-dimensional spaces. J. Nonparametr. Stat. **24**(2), 447–464 (2012)

[10] Ferraty, F., Vieu, P.: Nonparametric functional data analysis. Theory and practice. Springer, New York (2006)

[11] Gasser, T., Hall, P., Presnell, B.: Nonparametric estimation of the mode of a distribution of random curves. J. R. Stat. Soc. Ser. B Stat. Methodol. **60**(4), 681–691 (1998)

[12] Holm, S.: A simple sequentially rejective multiple test procedure. Scand. J. Statist. **6**(2), 65–70 (1979)

[13] Horváth, L., Kokoszka, P.: Inference for functional data with applications. Springer, New York (2012)

[14] Kawaguchi, A.: Estimating the correlation dimension from chaotic dynamical systems by U-statistics. Bulletin of Informatics and Cybernetics **34**(2), 143–150 (2002)

[15] Kokoszka, P., Reimherr, M.: Introduction to functional data analysis. CRC Press, Boca Raton, FL (2017)

[16] Lee, J.: U-statistics: Theory and Practice. Routledge (1990)

[17] Lehmann, E.L.: Elements of large-sample theory. Springer (1999)

[18] Li, W.V., Shao, Q.-M.: Gaussian processes: inequalities, small ball probabilities and applications. In: Stochastic Processes: Theory and Methods, pp. 533–597, Handbook of Statist. (vol. 19). North-Holland, Amsterdam (2001)

[19] Ramsay, J.O., Silverman, B.W.: Functional data analysis. Second edition. Springer, New York (2005)

Chapter 6
Principal Components Analysis of a Cyclostationary Random Function

Alain Boudou and Sylvie Viguier-Pla

Abstract Principal Components Analysis is a well-known method for reduction of dimension in Data Analysis. Considering a cyclostationary random function, we use appropriate transformations, based on spectral properties, in order to get a stationary random function, and then to process to a principal components analysis in the frequency domain. Then, a cyclostationary function is reconstituted as a summary of the initial cyclostationary function. Applications on simulated data illustrate the method.

6.1 Introduction

Cyclostationarity is a property which reveals some hidden periodicities in the energy flow of a signal. It is encountered in various phenomena, as in circular mechanisms, vibration and acoustic measurements (see Antoni [1] for a large class of examples). The approach of cyclostationary properties of time series may be applied for detection, motion analysis, monitoring, ... (see, for example, Lamraoui et al. [5], Zakaria [7]).

Many papers deal with cyclostationary signals, aiming at detecting a given phenomenum, or to estimate their shape. Most of them work in time domain, since the regularity of the shape is easy to write. Even if the registered signals are multidimensional, the studies often use univariate modeling, as in the examples cited above. As far as we know, there are very few studies considering multivariate cyclostationary functions (Bouleux et al. [4]).

Alain Boudou
Equipe de Stat. et Proba., Institut de Mathématiques, UMR5219, Université Paul Sabatier, 118 Route de Narbonne, F-31062 Toulouse Cedex 9, France, e-mail: boudou@math.univ-toulouse.fr

Sylvie Viguier-Pla (✉)
Equipe de Stat. et Proba., Institut de Mathématiques, UMR5219, Université Paul Sabatier, 118 Route de Narbonne, F-31062 Toulouse Cedex 9, France,
Université de Perpignan via Domitia, LAMPS, 52 av. Paul Alduy, 66860 Perpignan Cedex 9, France, e-mail: viguier@univ-perp.fr

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*, Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_6

Our work aims to consider the multivariate context, where the reduction of dimension is useful, in order to split the signal into main trend and particular skills, and in order to reduce the space occupied by the signal. Since time processes are particular functions, and since the set of time indexes can be replaced by other indices sets, we can classify this work in the field of the multidimensional functional analysis.

A multidimensional series $(X_n)_{n\in\mathbb{Z}}$ is said to be stationary when $E X_n{}^t\overline{X_m} = E X_{n-m}{}^t\overline{X_0}$, for any pair (n, m) of elements of $\mathbb{Z}$. In the particular case where the spectrum of $(X_n)_{n\in\mathbb{Z}}$ is concentrated on a finite number of elements of $[-\pi;\pi[$, it can be writen $X_n = \sum_{j\in J} e^{i\lambda_j n} Z_j$, where the Z_j's are such that $E Z_j{}^t\overline{Z'_j} = 0$, when $j \neq j'$, and $\lambda_j \in [-\pi;\pi[$.

The Principal Components Analysis (PCA) of $(X_n)_{n\in\mathbb{Z}}$ in the frequency domain consists in performing the PCA of each of the random vectors Z_j. This PCA gives best results, in terms of percent of retrieved variance, than PCA of the vectors X_n.

When a series $(X_n)_{n\in\mathbb{Z}}$ is $p-$cyclostationary, that is when $\operatorname{cov}(X_n, X_m) = \operatorname{cov}(X_{n+p}, X_{m+p})$, the $p-$multidimensional series $(Y_n)_{n\in\mathbb{Z}}$, where $Y_n = \begin{pmatrix} X_{np} \\ \vdots \\ X_{np+p-1} \end{pmatrix}$, is such that $E(Y_n^t\overline{Y_m}) = E(Y_{n-m}{}^t\overline{Y_0})$, and we can perform its PCA in the frequency domain. This is what we propose to expose in this text, for a $\Delta-$cyclostationary random function $(X_t)_{t\in\mathbb{R}}$, that is when $\operatorname{cov}(X_t, X_{t'}) = \operatorname{cov}(X_{t+\Delta}, X_{t'+\Delta})$. For this, using the previous model, to the Hilbert space $\mathbb{C}^p$, we substitute $L^2([0;\Delta[)$. More precisally, instead of performing PCA in the frequency domain of the series of vectors $((X_{np+k})_{k=0,\ldots,p-1})_{n\in\mathbb{Z}}$, we perform the PCA of the series of functions $((X_{n\Delta+t})_{t\in[0;\Delta[})_{n\in\mathbb{Z}}$.

6.2 Prerequisites and Notation

We denote by $\mathcal{B}$ the Borel $\sigma-$field of $\Pi = [-\pi;\pi[$. All the $\mathbb{C}-$Hilbert spaces (in particular H and H') of this text are supposed to be separable. We use the complex field because we need the Fourier transform. Our reference probability space is $(\Omega, \mathcal{A}, P)$. We define a probability measure η on ξ, $\sigma-$field of the subsets of $[0;1[$.

For any probability space $(F, \mathcal{F}, \mu)$, $\mathcal{L}^2_H(F, \mathcal{F}, \mu)$ stands for the set of measurable applications which square norm is $\mu-$integrable, from F in H, and $L^2_H(F, \mathcal{F}, \mu)$ stands for the set of the cosets of elements of $\mathcal{L}^2_H(F, \mathcal{F}, \mu)$. The index H is omited when $H = \mathbb{C}$. For any $\mathbb{C}-$Hilbert spaces H and H', we denote by $\sigma_2(H, H')$ the set, which is also a Hilbert space, of the Hilbert-Schmidt operators from H into H'.

Definition 1

A random measure (r.m.) Z, taking values in the $\mathbb{C}-$Hilbert space H, is a vector measure defined on $\mathcal{B}$ such that, for any pair (A, B) of disjoint elements of $\mathcal{B}$, $< ZA, ZB >= 0$.

Let us remark that the qualification "random" takes sense when $H = L^2(\Omega, \mathcal{A}, P)$.

Proposition 1 *If Z is a r.m. taking values in the $\mathbb{C}-$Hilbert space H, then the application $\mu_Z : A \in \mathcal{B} \mapsto \|ZA\|^2 \in \mathbb{R}_+$ is a bounded measure.*

We can now define the integral with respect to the r.m. as follows.

Proposition 2 *There exists an isometry, and only one, from $L^2(\Pi, \mathcal{B}, \mu_Z)$ onto $\overline{vect}\{ZA; A \in \mathcal{B}\}$, such that ZA is the image of 1_A, for any A of $\mathcal{B}$.*

The image of an element φ of $L^2(\Pi, \mathcal{B}, \mu_Z)$ by this isometry is denoted $\int \varphi \mathrm{d}Z$, and is named stochastic integral of φ with respect to the r.m. Z.

Let us now examine the notion of stationary series of elements of H.

Definition 2

A stationary series $(X_n)_{n\in\mathbb{Z}}$ of elements of H, is a family such that $< X_n, X_m >=< X_{n-m}, X_0 >$, for any pair (n, m) of elements of $\mathbb{Z}$.

Of course, when $H = L^2(\Omega, \mathcal{A}, P)$, and when $EX_n = 0$, we get the usual definition of the stationarity, because cov $(X_n, X_m) =< X_n, X_m >$.

There is a biunivoque relation between stationary series and r.m..

Proposition 3 *With any stationary series $(X_n)_{n\in\mathbb{Z}}$, of elements of H, we can associate a r.m. Z, and only one, taking values in H, such that $X_n = \int e^{i.n}dZ$, for any n of $\mathbb{Z}$.*

When U is a unitary operator (u.o.) of H, we have the following property.

Proposition 4 *For any X of H, $(U^n X)_{n\in\mathbb{Z}}$ is a stationary series.*

6.3 Principal Components Analysis in the Frequency Domain

This kind of analysis has first been studied by D. Brillinger [2]. His works have been completed and generalized by Boudou and Dauxois [3]. The aim of this analysis is to summarize a multidimensional stationary series by a series of lower dimension. Processing in the frequency domain enables to get rid of the problem of time dependence.

If X and Y are respectively elements of $L^2_H(\mathcal{A})$ and of $L^2_{H'}(\mathcal{A})$, the application $Y \otimes X : \omega \in \Omega \mapsto Y(\omega) \otimes X(\omega) \in \sigma_2(H', H)$ is measurable and of $P-$integrable norm. So the operator $\int Y \otimes X\mathrm{d}P$ is an Hilbert-Schmidt operator from H' into H.

Definition 1

A series $(X_n)_{n\in\mathbb{Z}}$ of elements of $L^2_H(\mathcal{A})$ is $H-$stationary when, for any pair (n, m) of elements of $\mathbb{Z}$, we have $\int X_m \otimes X_n dP = \int X_0 \otimes X_{n-m} dP$.

Let us remark that the $H-$stationarity implies the stationarity, but the converse is false.

Definition 2

Two series $(X_n)_{n\in\mathbb{Z}}$ and $(X'_n)_{n\in\mathbb{Z}}$, respectively $H-$stationary and $H'-$stationary, are stationarily correlated when, for any pair (n, m) of elements of $\mathbb{Z}$, we have $\int X_m \otimes X'_n dP = \int X_0 \otimes X'_{n-m} dP$.

Let then $(X_n)_{n\in\mathbb{Z}}$ be a $H-$stationary series. Our aim is to summarize it into a $\mathbb{C}^p-$stationary series $(X'_n)_{n\in\mathbb{Z}}$, which we want to be stationarily correlated with $(X_n)_{n\in\mathbb{Z}}$. In order to quantify the quality of the $p-$dimensional summary which is $(X'_n)_{n\in\mathbb{Z}}$, we will consider the series $(X''_n)_{n\in\mathbb{Z}} = (P_{\overline{\text{vect}}\{K\circ X'_n;(n,K)\in\mathbb{Z}\times\sigma_2(\mathbb{C}^p,H)\}} X_n)_{n\in\mathbb{Z}}$, where $P_{\overline{\mathrm{vect}}\{K\circ X'_n;(n,K)\in\mathbb{Z}\times\sigma_2(\mathbb{C}^p,H)\}}$ is the projector from $L^2_H(\mathcal{A})$ onto $\overline{\mathrm{vect}}\{K\circ X'_n;(n,K)\in\mathbb{Z}\times\sigma_2(\mathbb{C}^p,H)\}$. The series $(X''_n)_{n\in\mathbb{Z}}$ is $H-$stationary and stationarily correlated with $(X_n)_{n\in\mathbb{Z}}$, so the quantity $\|X_n - X''_n\| = \|X_0 - X''_0\|$ measures the quality of the summary.

Definition 3

The PCA in the frequency domain of a $H-$stationary series $(X_n)_{n\in\mathbb{Z}}$ is the search of a $\mathbb{C}^p-$stationary series $(X'_n)_{n\in\mathbb{Z}}$, stationarily correlated with $(X_n)_{n\in\mathbb{Z}}$, such that $\|X_0 - P_{\overline{vect}\{K\circ X'_n;(n,K)\in\mathbb{Z}\times\sigma_2(\mathbb{C}^p,H)\}}X_0\|$ is the smallest one.

For a same number of steps of PCA, this analysis produces smaller errors than the PCA of the random vectors X_n. As the $H-$stationary series $(X_n)_{n\in\mathbb{Z}}$ is stationary, there exists a r.m. Z, taking values in $L^2_H(\mathcal{A})$, such that $X_n = \int \mathrm{e}^{i.n}\mathrm{d}Z$, for any n of $\mathbb{Z}$. This r.m. is fundamental for the process of the PCA.

If $\{Z_j; j\in J\}$ is a finite family of elements of $L^2_H(\mathcal{A})$ such that $\int Z_j\otimes Z_{j'}\mathrm{d}P = 0$, for any pair (j,j') of distinct elements of J, and if $\{\lambda_j; j\in J\}$ is a family of distinct elements of Π, then $(\sum_{j\in J}\mathrm{e}^{i\lambda_j n}Z_j)_{n\in\mathbb{Z}}$ is a $H-$stationary series, of associated r.m. $Z = \sum_{j\in J}\delta_{\lambda_j}Z_j$, where δ_{λ_j} is the Dirac measure concentrated on λ_j. The PCA in the frequency domain is processed by the PCA of each of the vectors Z_j. This explains the name of PCA in the frequency domain.

6.4 Relation between the Spaces $L^2_{L^2(\mathcal{A})}(\xi)$ and $L^2_{L^2(\xi)}(\mathcal{A})$

Most of the results that we will recall here can be found in Schaeffer [6].

For any (y,h) of $L^2(\mathcal{A})\times L^2(\xi)$, we name yh (resp. hy) the application $\omega\in\Omega\mapsto y(\omega)h\in L^2(\xi)$ (resp. $t\in[0;1[\mapsto h(t)y\in L^2(\mathcal{A})$). The following property comes from the fact that $\overline{\mathrm{vect}}\{yh;(y,h)\in L^2(\mathcal{A})\times L^2(\xi)\} = L^2_{L^2(\xi)}(\mathcal{A})$, that $\overline{\mathrm{vect}}\{hy;(y,h)\in L^2(\mathcal{A})\times L^2(\xi)\} = L^2_{L^2(\mathcal{A})}(\xi)$, and that $< yh, y'h' >=< hy, y' >$, for any pair $((y,h),(y',h'))$ of elements of $L^2(\mathcal{A})\times L^2(\xi)$.

Proposition 1 *There exists an isometry $\mathcal{I}$, and only one, from $L^2_{L^2(\mathcal{A})}(\xi)$ on $L^2_{L^2(\xi)}(\mathcal{A})$, such that $\mathcal{I}(hy) = yh$, for any (y,h) of $L^2(\mathcal{A})\times L^2(\xi)$.*

The spaces $L^2_{L^2(\mathcal{A})}(\xi)$ and $L^2_{L^2(\xi)}(\mathcal{A})$ are isometric.

Proposition 2 *i) If X is an element of $L^2_{L^2(\xi)}(\mathcal{A})$, then the application $\widetilde{X} : y\in L^2(\mathcal{A})\mapsto \int yXdP\in L^2(\xi)$ is an Hilbert-Schmidt operator.*
ii) The application $X\in L^2_{L^2(\xi)}(\mathcal{A})\mapsto\widetilde{X}\in\sigma_2(L^2(\mathcal{A}),L^2(\xi))$ is an isometry.

So a series $(X_n)_{n\in\mathbb{Z}}$ of elements of $L^2_{L^2(\xi)}(\mathcal{A})$ is $L^2(\xi)-$stationary as soon as $\widetilde{X_n}\widetilde{X_m}^* = \widetilde{X_{n-m}}\widetilde{X_0}^*$ (since $\widetilde{X_n}\widetilde{X_m}^* = \int X_m\otimes X_n\mathrm{d}P = \int X_0\otimes X_{n-m}\mathrm{d}P = \widetilde{X_{n-m}}\widetilde{X_0}^*$), for any pair (n,m) of elements of $\mathbb{Z}$.

Exchanging the roles of $(\Omega,\mathcal{A},P)$ and $([0;1[,\xi,\eta)$, we also have the following results.

Proposition 3 *i) If Y is an element of $L^2_{L^2(\mathcal{A})}(\xi)$, then the application $\widetilde{Y} : h \in L^2(\xi) \mapsto \int hYd\eta \in L^2(\xi)$ is an Hilbert-Schmidt operator;*
ii) the application $Y \in L^2_{L^2(\mathcal{A})}(\xi) \mapsto \widetilde{Y} \in \sigma_2(L^2(\xi), L^2(\mathcal{A}))$ is an isometry.

Let us denote by γ (resp. Γ) the involutive antilinear bijection $y \in L^2(\mathcal{A}) \mapsto \overline{y} \in L^2(\mathcal{A})$ (resp. $h \in L^2(\xi) \mapsto \overline{h} \in L^2(\xi)$). For any (h, y) of $L^2(\mathcal{A}) \times L^2(\xi)$, we have $\gamma \circ \widetilde{hy} \circ \Gamma = \gamma \circ (\Gamma h \otimes y) \circ \Gamma = h \otimes \gamma y = \widetilde{yh}^*$. This can be generalized in the following way.

Proposition 4 *For any Y of $L^2_{L^2(\mathcal{A})}(\xi)$, we have $\gamma \circ \widetilde{Y} \circ \Gamma = \widetilde{\mathcal{I}Y}^*$.*

This last property implies the following one.

Proposition 5 *For any pair (Y, Y') of elements of $L^2_{L^2(\mathcal{A})}(\xi)$, we have $\widetilde{\mathcal{I}Y} \circ \widetilde{\mathcal{I}Y'}^* = \Gamma \circ \widetilde{Y}^* \circ \widetilde{Y'} \circ \Gamma$.*

With this last property, we will be able to prove the $L^2(\xi)-$stationarity of a series, in Section 6.6.

6.5 Cyclostationarity

Usually, a cyclostationary random function (r.f.) is a family $\{X_t; t \in \mathbb{R}\}$ of elements of $L^2(\mathcal{A})$ such that $\text{cov}(X_t, X_{t'}) = \text{cov}(X_{t+\Delta}, X_{t'+\Delta})$, Δ being an element of $\mathbb{R}_+^*$. In order to simplify notation, we will assume that $\Delta = 1$, without lost of generality, as we always can come back to the general case by a linear transformation.

Definition 1

A cyclostationary random function is a family $\{X_t; t \in \mathbb{R}\}$ of elements of $L^2(\Omega, \mathcal{A}, P)$ such that
i) the application $t \in [0; 1[\mapsto X_t \in L^2(\mathcal{A})$ is measurable and of $\eta-$integrable square norm;
ii) $< X_t, X_{t'} >=< X_{t+1}, X_{t'+1} >$, for any pair $(t, t')$of elements of $\mathbb{R}$.

Point i) of this definition is very little restrictive, with respect to the usual definition. It is satisfied as soon as the $\sigma-$field ξ is the trace of $\mathcal{B}_\mathbb{R}$ on $[0; 1[$, and as the application $t \in \mathbb{R} \mapsto X_t \in L^2(\mathcal{A})$ is continuous.

If $(X_t)_{t\in\mathbb{R}}$ is a stationary continuous r.f., that is when $t \in \mathbb{R} \mapsto X_t \in L^2(\mathcal{A})$ is continuous and when $< X_t, X_{t'} >=< X_{t-t'}, X_0 >$, for any (t, t') of $\mathbb{R}^2$, then it is also a cyclostationary r.f..

6.6 Definition of a $L^2(\xi)-$stationary Series from a Cyclostationary r. f.

Let then $(X_t)_{t\in\mathbb{R}}$ be a cyclostationary r.f..

Proposition 1 *There exists a u.o. V of $L^2(\mathcal{A})$ such that $VX_t = X_{t+1}$, for any t of $\mathbb{R}$.*

So we can prove that $(V^n X_t)_{n\in\mathbb{Z}} = (X_{t+n})_{n\in\mathbb{Z}}$, for any t of $[0; 1[$. We can then enunciate the following.

Proposition 2 *i) for any n of $\mathbb{Z}$, the application $t \in [0;1[\mapsto X_{t+n} \in L^2(\mathcal{A})$ is a representative of an element Y_n of $L^2_{L^2(\mathcal{A})}(\xi)$;*
ii) the series $(Y_n)_{n\in\mathbb{Z}}$ of elements of $L^2_{L^2(\mathcal{A})}(\xi)$ is stationary;
iii) $(\mathcal{I}Y_n)_{n\in\mathbb{Z}}$ is a $L^2(\xi)-$stationary series of elements of $L^2_{L^2(\xi)}(\mathcal{A})$.

Let us give some elements for a proof of this property.

We consider the application $\widehat{V}$ which transforms the coset of the element X of $\mathcal{L}^2_{L^2(\mathcal{A})}(\xi)$ into the coset of $V \circ X$. The operator $\widehat{V}$ is a u.o. of $L^2_{L^2(\mathcal{A})}(\xi)$ such that $\widehat{V}^n Y_0 = Y_n$. Indeed, $\widehat{V}^n Y_0$ is the coset of the application $V^n \circ (t \in [0;1[\mapsto X_t \in L^2(\mathcal{A}))$, so of $(t \in [0;1[\mapsto V^n X_t \in L^2(\mathcal{A}))$, or evenmore of $t \in [0;1[\mapsto X_{t+n} \in L^2(\mathcal{A})$. Points i) and ii) are then proved. We can prove that $\widetilde{Y_n} = V^n \circ \widetilde{Y_0}$. The last point comes from Proposition 6.4.5: for any pair (n,m) of elements of $\mathbb{Z}$,

$$\widetilde{\mathcal{I}Y_n}\,\widetilde{\mathcal{I}Y_m}^{\,*} = \Gamma \circ \widetilde{Y_n}^{\,*}\widetilde{Y_m} \circ \Gamma = \Gamma \circ \widetilde{Y_0}^{\,*} \circ V^{-n} \circ V^m \circ \widetilde{Y_0} \circ \Gamma = \widetilde{\mathcal{I}Y_{n-m}}\,\widetilde{\mathcal{I}Y_0}^{\,*}.$$

We can now define the series associated with a cyclostationary r.f..

Definition 1

The series $(\mathcal{I}Y_n)_{n\in\mathbb{Z}}$ is named $L^2(\xi)-$stationary series deduced from the cyclostationary r.f. $(X_t)_{t\in\mathbb{R}}$.

Of course, it is possible to process a PCA in the frequency domain of the $L^2(\xi)-$stationary series $(\mathcal{I}Y_n)_{n\in\mathbb{Z}}$.

The p first steps will give a $\mathbb{C}^p-$stationary series $(X'_n)_{n\in\mathbb{Z}}$. As for the reconstruction of the series, it will give a $L^2(\xi)-$stationary series $(X''_n)_{n\in\mathbb{Z}}$, stationarily correlated with $(\mathcal{I}Y_n)_{n\in\mathbb{Z}}$. Then it is possible to define a cyclostationary r.f. $(\mathcal{X}_t)_{t\in\mathbb{R}}$ for which the deduced $L^2(\xi)-$stationary series is $(X''_n)_{n\in\mathbb{Z}}$. So we can write

$$\int \|\mathcal{X}_{t+n} - X_{t+n}\|^2 \mathrm{d}\eta(t) = \|X''_n - \mathcal{I}Y_n\|^2 = \|X''_0 - \mathcal{I}Y_0\|^2,$$

for any n of $\mathbb{Z}$.

The cyclostationary r.f. $(\mathcal{X}_t)_{t\in\mathbb{R}}$ is a reconstitution of the data.

Let us remark that, as a stationary continuous r.f. is a cyclostationary r.f., we can process to such a PCA for it.

Let us now examine the respective r.m.'s associated with the stationary series $(\mathcal{I}Y_n)_{n\in\mathbb{Z}}$ and $(Y_n)_{n\in\mathbb{Z}}$. We denote by Z_Y the r.m. associated with this last, and, for any t of $[0;1[$, by Z_t the r.m. associated with $(X_{t+n})_{n\in\mathbb{Z}}$, which is stationary, as $(X_{t+n})_{n\in\mathbb{Z}} = (V^n X_t)_{n\in\mathbb{Z}}$. So we have the following property.

Proposition 3 *For any A of $\mathcal{B}$, $Z_Y A$ is the coset of the application $t \in [0;1[\mapsto Z_t A \in L^2(\mathcal{A})$.*

As for the r.m. associated with the $L^2(\xi)-$stationary series $(\mathcal{I}Y_n)_{n\in\mathbb{Z}}$, which is important for PCA is the frequency domain, it is $\mathcal{I} \circ Z_Y$.

All these results can be generalized to other types of cyclostationary functions $(X_g)_{g\in G}$, where G is a locally compact abelian group.

6.7 Simulation of a Cyclostationary r. f. and PCA

Let us examine the analysis that we just have recommended, for a simulated cyclostationary function. For graphical representation facilities, we chose a simulation which gives a real valued cyclostationary function. For this, let us consider $\Omega = \{\omega_1, \ldots, \omega_k\}$, $k > 1$, and $P(\{\omega_j\}) = \frac{1}{k}$. All the matrix expressions will be relatively to the orthonormal basis $\{y_1, \ldots, y_k\}$, where $y_j = \sqrt{k} 1_{\{\omega_j\}}$.

Let V be a u.o. of $L^2(\mathcal{A})$, of matrix B (where $B = \overline{B}$), and $\{f_1, \ldots, f_k\}$ be a family of elements of $L^2([0; 1[)$ $(f_j = \overline{f_j})$.

Then we build a cyclostationary function $\{X_t, t \in \mathbb{R}\}$ as follows:

$$X_t = V^{[t]}(\textstyle\sum_{j=1}^k f_j(t - [t]) y_j) = \sum_{j=1}^k f_j(t - [t]) V^{[t]} y_j,$$

where $[t]$ stands for the integer part of t.

The matrix expression of what precedes is

$$\frac{1}{\sqrt{k}} \begin{pmatrix} X_t(\omega_1) \\ \vdots \\ X_t(\omega_k) \end{pmatrix} = \sum_{j=1}^{k} f_j(t - [t])(B^{[t]})_{.j}.$$

So we can get the trajectories $X_t(\omega_l) = \sum_{j=1}^k f_j(t - [t])(B^{[t]})_{lj}$.

Figure 6.1 shows some trajectories when $k = 10$ and $f_j(t) = t^j$.

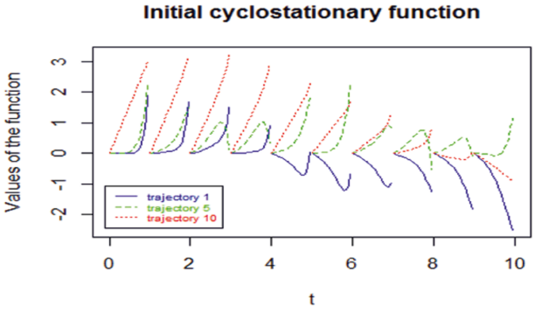


Fig. 6.1 Trajectories 1, 5 and 10 of the simulated cyclostationary function X_t

.

If $V = \sum_{l \in L} e^{i\mu_l} P_l$ is the spectral decomposition of V, then the deduced $L^2(\xi)$−stationary series $(\mathcal{I}Y_n)_{n \in \mathbb{Z}}$ is such that $\mathcal{I}Y_n = \sum_{l \in L} e^{i\mu_l n} Z_l$, where $Z_l = \sum_{j=1}^k (P_l y_j) f_j$.

The matrix associated with $\widetilde{Z}_l^* \widetilde{Z}_l$ is $\overline{M_l} T \overline{M_l}$, where M_l is the matrix associated with the P_l projector, and T is such that $T_{uv} =< f_u, f_v >_{L^2([0;1[)}$.

The first q steps of the PCA of $(\mathcal{I}Y_n)_{n \in \mathbb{Z}}$ give the following reconstitution:

$$\frac{1}{\sqrt{k}} \begin{pmatrix} X_t(\omega_1) \\ \vdots \\ X_t(\omega_k) \end{pmatrix} = \sum_{j=1}^{k} \sum_{l \in L} \sum_{u=1}^{k} ({}^t M_l \sum_{v=1}^{q} U_v^l \, {}^t\overline{U_v^l})_{uj} f_u(t - [t])(B^{[t]})_{.j},$$

where U_v^l is the v^{th} normed eigenvector of the matrix $\overline{M_l} T \overline{M_l}$.

Figure 6.2 shows the reconstituted trajectories of Figure 6.1 when $q = 1$.

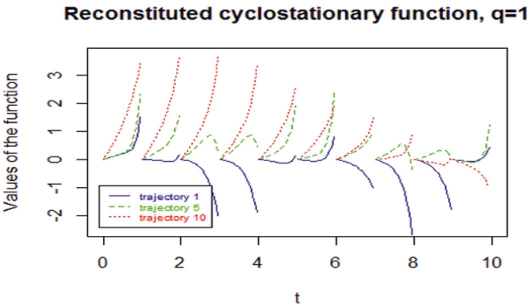


Fig. 6.2 One-dimensional reconstituted trajectories 1, 5 and 10 of the simulated cyclostationary function X_t

For this simulation, the sum of squares of the errors are respectively equal to 77.6, 5.1 and 0 when $q = 1$, $q = 2$ and $q = 3$. The choice of B such that at least one of the eigenvalue is triple and the other eigenvalues are of multiplicity less or equal to three makes a perfect reconstitution for $q = 3$. When all the eigenvalues are single, the first step of the PCA reconstitutes the all cyclostationary function.

References

[1] Antoni, J.: Cyclostationarity by examples. Mechanical Systems and Signal Processing **23**, 987–1036 (2009)

[2] Brillinger, D.M.: Time Series Data Analysis and Theory. Reprint of the 1981 edition. Classics in Applied Mathematics 36. Society for Industrial Applied Mathematics (SIAM), Philadelphia (2001)

[3] Boudou, A.,Dauxois, J.: Principal Component Analysis for a Stationary Random Function Defined on a Locally Compact Abelian Group. J. Multivariate Anal. **51**, 1–16 (1994)

[4] Bouleux, G., Dugast, M., Marcon, F.: Information Topological Characterization of Periodically Correlated Processes by Dilation Operators. IEEE Trans. on Information Theory **65**(10), 6484–6495 (2019)

[5] Lamraoui, M., Thomas, M., El Badaoui, M.: Cyclostationarity approach for monitoring chatter and tool wear in high speed milling. Mechanical Systems and Signal Processing **44**, 177–198 (2014)

[6] Schaeffer, H.H.: Topological Vector Spaces. Springer-Verlag, New York Heideberg Berlin (1971)

[7] Zakaria, F.: Analyse de la locomotion humaine: exploitation des propriétés de cyclostationaryté des signaux. PhD thesis, Université Jean Monnet, Saint Étienne, France (2015)

Chapter 7
Level Set and Density Estimation on Manifolds

Alejandro Cholaquidis, Ricardo Fraiman and Leonardo Moreno

Abstract Given an iid sample of a distribution supported on a smooth manifold $M \subset \mathbb{R}^d$, which is assumed to be absolutely continuous w.r.t the Hausdorff measure inherited from the ambient space, we tackle the problem of the estimation of the level sets of the density f. A consistent estimator in both Hausdorff distance and distance in measure is proposed. The estimator is the level set of the kernel-based estimator of the density f. We prove that the kernel-based density estimator converges uniformly to the unknown density f, the consistency of the level set and the consistency of the boundary of the level set estimator. The performance of our proposal is illustrated through some simulated examples.

7.1 Introduction

The estimation of level sets $L_f(\lambda) = \{x : f(x) \geq \lambda\}$, where f is an unknown density function on $\mathbb{R}^d$ and $\lambda > 0$ is a given constant, has been previously considered by several authors; see, for instance, [14], [22], [8], [18] [25], [27] for consistency results and rates of convergence, while the asymptotic distribution was derived in [5]. Some relevant applications are mode estimation [19], [22], clustering ([9], [10]) or detection of abnormal behaviour in a system ([12], [2], [1]).

However, this problem is less developed when the underlying density has its support on a Riemannian manifold. The statistical analysis of several problems when data takes values on a Riemannian manifold has received much attention in the last few years. One of the reasons is that at present, we are interested in the statistical analysis of more complex objects and structures. References on the subject are

Alejandro Cholaquidis
Universidad de la República, Uruguay, e-mail: acholaquidis@cmat.edu.uy

Ricardo Fraiman (✉)
Universidad de la República, Uruguay, e-mail: rfraiman@cmat.edu.uy

Leonardo Moreno
Universidad de la República, Uruguay, e-mail: mrleo@iesta.edu.uy

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_7

numerous. We refer to [17], [3], and [21] and the references therein for an overview. In the following, we address the problem of level set estimation in this setup.

This problem requires us to first tackle the estimation of the underlying density, a problem that has been addressed in the manifold framework, for instance, in [15] for a manifold without boundary. This manuscript aims to extend previous result to the case of manifolds with boundary and to obtain as a by-product the consistency (w.r.t the Hausdorff distance and the distance in measure) of the natural level set estimators, which are the level set of the density estimator. We also prove that the boundary of the level set of the density estimator is consistent in Hausdorff distance. Let us introduce more formally our problem.

Given a d'-dimensional Riemannian manifold $M \subset \mathbb{R}^d$, where $d' \leq d$ and d' is assumed to be known, the aim is to estimate the level sets

$$L_f(\lambda) = \{x \in M : f(x) \geq \lambda\}$$

of the density f of a random vector X with support M from an iid sample $X_1, \ldots, X_n$ with distribution f. First, we will consider the case where λ is such that $L_f(\lambda) \cap \partial M = \emptyset$, where ∂M denotes the boundary of M. Next, we will tackle the problem where $L_f(\lambda) \cap \partial M \neq \emptyset$. To do so, we will use the plug-in estimator $\hat{L}_{f_{n,h}}(\lambda) = \{x : f_{n,h}(x) \geq \lambda\}$, where $f_{n,h}$ is a kernel-based estimator with bandwidth $h = h_n \longrightarrow 0$. In the following, we assume that λ is fixed.

In Section 7.3, we prove that the kernel-based density estimator converge uniformly to the unknown density f. The consistency of the level sets in the Hausdorff distance and in measure is addressed in Section 7.4. Consistency in the Hausdorff metric of level sets under r–convexity is shown in Section 7.5, while in Section 7.6, we provide some simulation results.

7.2 Notation and Geometric Framework

If $B \subset \mathbb{R}^d$ is a Borel set, we denote by $|B|$ its Lebesgue measure and by $\overline{B}$ its closure. Given a set A on a topological space, the interior of A with respect to the underling topology is denoted by $\mathring{A}$. The k-dimensional closed ball of radius ε centred at x will be denoted by $\mathcal{B}_k(x, \varepsilon) \subset \mathbb{R}^d$ (when $k = d$ the index will be omitted), and its Lebesgue measure is denoted by $\sigma_k = |\mathcal{B}_k(x, 1)|$.

In the following, $M \subset \mathbb{R}^d$ is a compact d'-dimensional manifold of class C^2 (also called a d'-regular surface of class C^2). We consider the Riemannian metric on M inherited from $\mathbb{R}^d$. We denote by $\rho(x, y)$ the geodesic distance between x, y and given a set $A \subset M$, we denote $B_\rho(A, r) = \{x \in M : d(x, A) < r\}$. When M has a boundary, as a manifold, it is be denoted by ∂M. We denote for $\delta > 0$, $M_\delta : \{x \in M : \rho(x, \partial M) \geq \delta\}$ When M is orientable, it has a unique associated volume form ω such that $\omega(e_1, \ldots, e_{d'}) = 1$ for all oriented orthonormal bases $e_1, \ldots, e_{d'}$ of $T_x M$. Then, if $g : M \longrightarrow \mathbb{R}$ is a density function, we can define a new measure $\mu(B) = \int_B g d\omega$, where $B \subset M$ is a Borel set. Since we are only interested in measures, which can be defined even if the manifold is not orientable, although in a slightly less intuitive way, the orientability hypothesis is dropped in the following.

Given a point $x \in M$, b_x is the geodesic distance from x to the boundary ∂M of M, or is ∞ if $\partial M = \emptyset$. Given $x \in M$ and $f : M \longrightarrow \mathbb{R}$ we denote by $d_x f$ the differential of f at x.

Recall that given two non-empty compact sets $A, C \subset \mathbb{R}^d$, the Hausdorff distance between A and C is defined as

$$d_H(A, C) = \max\left\{\max_{a\in A}\rho(a, C),\ \max_{c\in C}\rho(c, A)\right\},\ \text{where } \rho(a, C) = \inf_{c\in C}\rho(a, c). \tag{7.1}$$

Given two Borel sets $A, B \subset M$, the distance in measure between them is $d_\mu(A, B) = \mu(A \setminus B) + \mu(B \setminus A)$.

7.3 Density Estimation

The aim of this section is to prove that the kernel-based density estimator proposed in [4], denoted by $f_{h,n}$, converges uniformly to the density f. For simplicity, we assume that K is the gaussian kernel, i.e, $K(\|x\|) = \pi^{-d'/2}\exp(-\|x\|^2)$. Let $h = h_n \longrightarrow 0$; then,

$$f_{h,n}(x) = \frac{1}{nm_0(x)h^{d'}}\sum_{i=1}^{n} K\Big(\frac{\|x - X_i\|}{h}\Big) \text{ where } m_0(x) = \pi^{-1/2}\int_{-\infty}^{b_x/h}\exp(-z^2)dz, \tag{7.2}$$

b_x is the geodesic distance from x to ∂M or is ∞ if $\partial M = \emptyset$. Equation (5) in [4] states that the bias of $f_{h,n}(x)$ is

$$E(f_{h,n}(x)) - f(x) = hm_1(x)\langle\eta_x, \nabla f(x)\rangle + O_x(h^2) \text{ where } m_1(x) = \frac{1}{2\sqrt{\pi}}\exp(-b_x^2/h^2). \tag{7.3}$$

First, we will tackle the case where the level λ is such that $L_f(\lambda) \cap \partial M \neq \emptyset$. In this case, if $x \in L_f$, then $m_0(x) \longrightarrow 1$, so we will replace the estimator (7.2) by

$$\hat{f}_{h,n}(x) = \frac{1}{nh^{d'}}\sum_{i=1}^{n} K\Big(\frac{\|x - X_i\|}{h}\Big).$$

7.3.1 First Case

Theorem 1 *Let M be a C^2 compact d'-dimensional submanifold of $\mathbb{R}^d$. Let X be a random vector with support M whose density f is assumed to be C^2. Let $h \longrightarrow 0$ and $\beta_n \longrightarrow \infty$ such that $\beta_n h^2 \longrightarrow 0$, $nh^{d'}/(\beta_n^2\log(n)) \longrightarrow \infty$; then,*

$$\beta_n \sup_{x\in M_0} |\hat{f}_{h,n}(x) - f(x)| \longrightarrow 0 \quad a.s.$$

for any closed subset $M_0 \subset M$ such that $\inf_{x\in M_0}\rho(x, \partial M) > 0$.

The following results are more general, so the theorem allows the compact set M_0 to depend on n. It is proven in the same manner as Theorem 1,

Theorem 2 *Let M be a C^2 compact d'-dimensional submanifold of $\mathbb{R}^d$. Let X be a random vector with support M whose density f is assumed to be C^2. Let $h \longrightarrow 0$ and $\beta_n \longrightarrow \infty$, such that $\beta_n h^2 \longrightarrow 0$, $nh^{d'}/(\beta_n^2 \log(n)) \longrightarrow \infty$; then,*

$$\beta_n \sup_{x \in M_n} |\hat{f}_{h,n}(x) - f(x)| \longrightarrow 0 \quad a.s.$$

for any sequence of closed subsets $M_n \subset M$ such that $\inf_{x \in M_n} \rho(x, \partial M)/h \longrightarrow \infty$.

7.4 Level Set Estimation

The estimation of the level sets of the density in Hausdorff distance and in measure when it does not meet the boundary of the manifold (in case it has) is proven in the following result.

Theorem 3 *Let M and f in the hypotheses of Theorem 2. Assume that the level $\lambda > 0$ fulfills that for all x such that $f(x) = \lambda$, there exists $a_n, b_n \longrightarrow x$ such that $f(a_n) > \lambda$ and $f(b_n) < \lambda$ and the boundary $\partial\{f \geq \lambda\}$ is non-empty. Then, with probability one,*

1 $d_H(\partial L_{\hat{f}_{n,h}}(\lambda), \partial L_f(\lambda)) \longrightarrow 0$;
2 $d_H(L_{\hat{f}_{n,h}}(\lambda), L_f(\lambda)) \longrightarrow 0$;
3 If, moreover, $d_x f \neq 0$ for all x such that $f(x) = \lambda$, $d_\mu(L_{\hat{f}_{n,h}}(\lambda), L_f(\lambda)) \longrightarrow 0$.

Theorem 4 *Let M and f be as in the hypotheses of Theorem 2. Assume that the level $\lambda > 0$ fulfills that for all x with $f(x) = \lambda$, there exists $a_j \longrightarrow x$, $a_j \in \mathring{M}$, such that $f(a_j) > \lambda$ for all j. Then,*

$$d_H\big(L_{\hat{f}_{n,h}}(\lambda), L_f(\lambda)\big) \longrightarrow 0, \quad a.s., \text{ as } n \longrightarrow \infty.$$

7.5 Manifold Level Set Estimation under r-convexity

In a Euclidean space, a set A is said to be r-convex (for some $r > 0$) if $A = C_r(A)$, where $C_r(A)$ is the r-convex hull of A, i.e. the intersection of the complements of all open balls of radii r that does not meet A. It is a natural generalization of convexity (the half spaces are replaced by balls), and it has been widely studied in set estimation literature (see, for instance, [27, 26] [23] and [20]). Additionally, as is pointed out in [23], this concept “is closely related to the notion of alpha-shapes that arises in the literature of computational geometry"; see [13]. Departing from the idea of r-convexity, several generalizations have been given (see, for instance, [6]). If the underlying space is not a Euclidean space but rather is any Riemannian manifold M endowed with the geodesic distance ρ, the natural generalization is to replace the Euclidean balls with geodesic balls. According to this idea, given $r > 0$, we will say that a set $A \subset M$ is r-convex if it is equal to its r-convex hull in M, i.e. the intersection of the complement of all open geodesic balls of radii r that does not meet A.

Theorem 5 *Under the hypotheses of Theorem 4, assume also that the level set $L_f(\lambda)$ is r-convex and $L_{\hat{f}_{n,h}}(\lambda)$ is r-convex a.s., for some $r > 0$. Then,*

$$d_H\big(C_r(\{X_i : \hat{f}_{h,n}(X_i) > \lambda\}), C_r(\{X_i : f(X_i) > \lambda\})\big) \to 0, \quad a.s.$$

and

$$d_H\big(C_r(\{X_i : \hat{f}_{h,n}(X_i) > \lambda\}), L_f(\lambda)\big) \to 0 \quad a.s.$$

7.6 Simulation Results

To assess the performance of our proposal, we will perform a simulation example with two scenarios. In the first one, we consider a distribution on the positive cone of covariance matrices, which is a three dimensional manifold when endowed with a Riemannian structure given below. In the second one, we will consider the torus with the metric inherited from R^3. In this case, we consider two distributions, the first one being unimodal and the last one being a mixture of distributions.

7.6.1 Positive-definite Matrices

Let us denote by $(\mathbb{P}_d, g)$ the set of positive-definite $d \times d$-covariance matrices. Given two matrices $A, B \in \mathbb{P}_d$, the geodesic curve joining A and B is

$$\gamma(s) = A^{1/2}(A^{-1/2}BA^{-1/2})^s A^{1/2} \quad \text{for all } s \in [0, 1].$$

The geodesic distance is given by $d_g(A, B) = \| \ln(A^{-1/2}BA^{-1/2})\|$, where $\| \cdot \|$ is the Hilbert–Schmidt norm.

We consider, for $d = 2$, the Wishart distribution $\mathcal{W}_2(\Sigma, m)$ on $\mathbb{P}_2$ with parameters $m = 10$ and $\Sigma = (1/2)I_2$. An easy way to obtain a matrix S with this distribution is to define $S = X_1X_1' + \cdots + X_mX_m'$, where $X_1, \ldots, X_m$ is an iid random sample of a multivariate Gaussian distribution with mean 0 and covariance matrix Σ.

As is well known, $(\mathbb{P}_2, g)$ can be represented as a cone in $\mathbb{R}^3$. In Figure 7.1, we show the projections of a sample of size 1000, drawn from a Wishart distribution with $m = 10$ and $\Sigma = (1/4)I$, together with the convex hull of the λ level set $L_{\mathcal{W}}(\lambda)$ (in blue) and the convex hull of the level set estimator $L_{\hat{\mathcal{W}}_{n,h}}(\lambda)$ (in red) for $\lambda = 0.06$ and $h = 0.1$. The estimator was obtained with a sample of size $n = 10000$. The Hausdorff distance between the level sets in $\mathbb{R}^3$ is 0.56. In Table 7.1, we report the mean over 500 replications of the Hausdorff distance (d_H) between both sets for different sample sizes $n \in \{1000, 5000, 10000, 20000\}$.

n	h	d_H
1000	0.20	0.732
5000	0.15	0.6
10000	0.10	0.56
20000	0.05	0.4

Table 7.1 Hausdorff distance between the true level set $L_{\mathcal{W}}(\lambda)$ and the estimator $L_{\hat{\mathcal{W}}_{n,h}}(\lambda)$ for $\lambda = 0.5$ and different values of h.

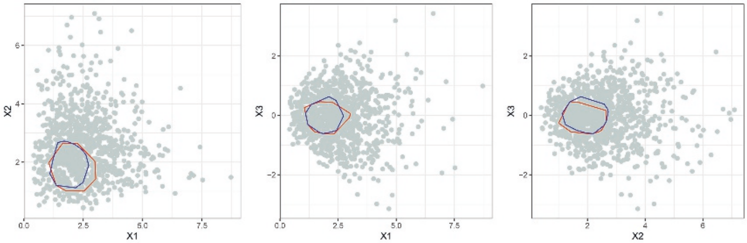


Fig. 7.1 Projections of a sample of size 1000 drawn from a Wishart distribution with $m = 10$ and $\Sigma = (1/4)I$, together with the convex hull of the λ level set (in blue) and the convex hull of the level set estimator (in red), for $\lambda = 0.06$ and $h = 0.1$.

7.6.2 The Torus

In the torus $\mathbb{T}^2 = S^1 \times S^1$, we consider the multivariate von Mises distribution, denoted by $\mathcal{MVM}(\mu, \kappa, \Delta)$. The density at $\theta \in \mathbb{T}$ is given by

$$f(\theta; \mu, \kappa, \Delta) = \frac{1}{Z(\kappa, \Delta)} \exp\{\kappa^\top c(\theta) + s(\theta)\Delta s(\theta)/2\},$$

where $\mu \in \mathbb{T}^2$ (this parameter is called mean), $\kappa \geq 0 \in \mathbb{R}^d$ (concentration parameter), $\Delta = (\lambda_{i,j})$ is a symmetric matrix on $\mathbb{R}^{d \times d}$ with null diagonal entries ($\lambda_{i,i} = 0$ for all $i \in \{1, \ldots, d\}$), and $Z(\kappa, \Delta)$ is a normalization constant. The functions c_i and s_i are defined by $c_i(\theta) = \cos(\theta_i - \mu_i)$ and $s_i(\theta) = \sin(\theta_i - \mu_i)$ for all $i \in \{1, \ldots, d\}$. In Figure 7.2 (left panel), we show (in yellow) a sample of size 2000 from a $\mathcal{MVM}_1(\mu_1, \kappa_1, \Delta_1)$ distribution with

$$\mu_1 = (\pi/2, 0), \quad \kappa_1 = (20, 20), \quad \Delta_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{7.4}$$

In the right panel of Figure 7.2, we show (in yellow) a sample of size 2000 from a mixture law given by

$$0.4\mathcal{MVM}_1(\mu_2, \kappa_1, \Delta_1) + 0.6\mathcal{MVM}_2(\mu_3, \kappa_1, \Delta_1), \tag{7.5}$$

where $\mu_2 = (\pi/2, 0)$ y $\mu_3 = (\pi/2, \pi/4)$. In all cases, we consider $\lambda = 0.8$ and bandwidth $h = 0.2$. The boundary of the theoretical level set is shown in red, while the boundary of the estimator is shown in magenta.

The Hausdorff distances between the theoretical curve and the estimated one are 0.066 and 0.107.

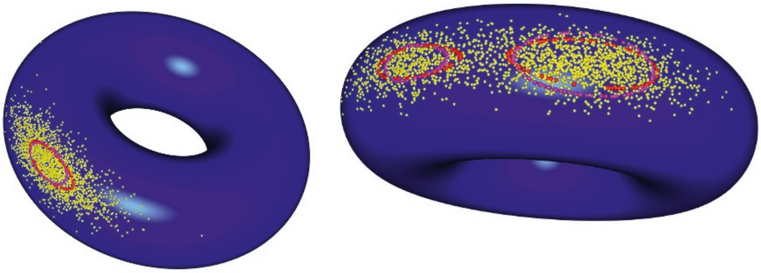


Fig. 7.2 Left panel: a sample of size 2000 from a $\mathcal{MVM}_1(\mu_1, \kappa_1, \Delta_1)$ distribution with μ_1, κ_1 and Δ_1 given in (7.4). Right panel: a sample of size 2000 from the mixture law given in (7.5). In both cases, the data are shown in yellow, whereas the boundary of the true level sets is shown (in red) together with the estimated boundary (in magenta).

7.6.3 The Sphere

Finally, we considered the sphere $S^2 \subset \mathbb{R}^3$ endowed with the Riemannian metric inherited from $\mathbb{R}^3$. The sample is drawn from a the mixture of two von Mises–Fisher distributions given by

$$f(x, \mu, \kappa) = C(x)e^{\kappa\mu^\top x}\mathbb{I}_{S_2}(x),$$

where $\kappa \geq 0$ and $\mu \in S^2$ are the concentration and directional mean parameters, respectively. $C(x)$ is the normalizing constant; see [17].

The mixture is given by,

$$0.5f(\cdot, (-1, -1/4, 0), 40) + 0.5f(\cdot, (-1, 1/4, 0), 40)\,. \tag{7.6}$$

In Figure 7.3, we show (left panel) a sample of size $n = 500$ on S^2, together with the estimated level set (in red) and the true level set (in blue). In the right panel, we show the stereographic projections of the sample and the estimators. The Hausdorff distance between the theoretical curve and the estimated (on the stereographic projections) curve is 0.018.

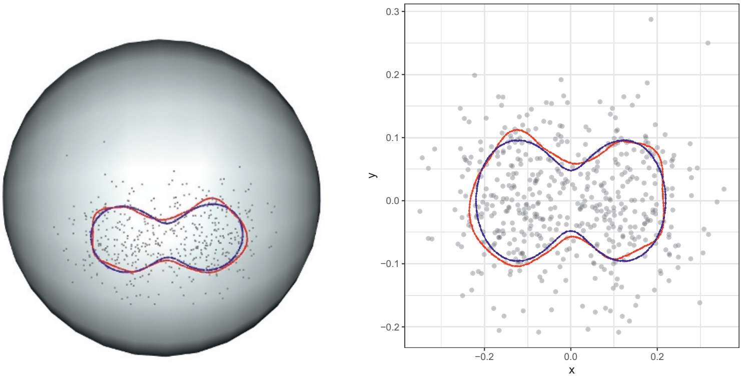


Fig. 7.3 Left panel: A sample of size 500 from the mixture of two von Mises–Fisher distributions given in equation (7.6). Right panel: the stereographic projections of the sample and the level sets. In both cases, the estimator is shown in red, while the true underlying level set is shown in blue.

References

[1] Baillo, A.L: Total error in a plug-in estimator of level sets. Statist. Probab. Lett. **65**, 441–417 (2003)

[2] Baillo, A., Cuesta-Albertos, J.A., Cuevas, A.: Convergence rates in nonparametric estimation of level sets. Statist. Probab. Lett. **53**, 27–35 (2001)

[3] Bhattacharya, A., Bhattacharya, R.: Nonparametric inference on manifolds: with applications to shape spaces. Vol. 2. Cambridge University Press (2012)

[4] Berry, T., Sauer, T.: Density estimation on manifolds with boundary. Computational Statistics & Data Analysis **107**, 1–17 (2017)

[5] Chen, Y.-C., Genovese, C.R., Wasserman, L.: Density Level Sets: Asymptotics, Inference, and Visualization. J. Amer. Statist. Assoc. **112**, 1684–1696 (2017)

[6] Cholaquidis, A., Cuevas, A., Fraiman, R.: On Poincaré cone property. Ann. Statist. **42**, 255–284 (2014)

[7] Cuevas, A., González-Manteiga, W., Rodríguez-Casal, A.: Plug-in estimation of general level sets. Aust. N. Z. J. Stat. **48**(1), 7–19 (2006)

[8] Cuevas, A., Fraiman, R.: A plug-in approach to suppoprt estimation. Ann. Statist. **25**, 2300–2312 (1997)

[9] Cuevas, A., Febrero, M. Fraiman, R.: Estimating the number of clusters. Canad. J. Statist. **28**, 367–382 (2000)

[10] Cuevas, A., Febrero, M., Fraiman, R.: Cluster analysis: a further approach based on density estimation. Comput. Statist. Data Anal. **36**, 441–459 (2000)

[11] Cuevas, A., Fraiman, R., Pateiro-López, B.: On statistical properties of sets fulfilling rolling-type conditions. Adv. in Appl. Probab. **44**, 311–329 (2012)

[12] Devroye, L., Wise, G.L.: Detection of abnomral behaviour via nonparametric estimation of the support. SIAM J. Appl. Math. **38**, 480–488 (1980)

[13] Edelsbrunner, H., Mücke, E.P.: Three-dimensional alpha-shapes. ACM Transactions on Graphics **13**, 43–72 (1994)

[14] Hartigan, J.A.: Estimation of convex density contour in two dimensions. J. Amer. Statist. Assoc. **82**, 267–270 (1987)

[15] Henry, G., Rodriguez, D.: Kernel density estimation on Riemannian manifolds: Asymptotic results. Journal of Mathematical Imaging and Vision **34**(3), 235–239 (2009)

[16] Kim, J., Shin, J., Rinaldo, A., Wasserman, L.: Uniform Convergence of the Kernel Density Estimator Adaptive to Intrinsic Volume Dimension. arXiv: 1810.05935v2 (2019)

[17] Mardia K.V.: Statistics of directional data. Academic Press (1972)

[18] Molchanov, I.: A limit theorem for solutions of inequalities. Scandinavian Journal of Statistics **25**, 235–242 (1998)

[19] Müller, D.W., Sawitzki, G.: Excess mass estimates and test of multimodality. J. Amer. Staitist. Assoc. **86**, 738–746 (1991).

[20] Pateiro-López, B., Rodríguez-Casal, A.: Length and surface area estimation under smoothness restrictions. Adv. in Appl. Probab. **40** 348–358 (2008)

[21] Patrangenaru, V., Ellingson, L.: Nonparametric statistics on manifolds and their applications to object data analysis. CRC Press (2015)

[22] Polonik, W.: Measuring mass concentration and estimating density contour clusters - an excess mass approach. Ann. Statist. **23**, 855–881 (1995)

[23] Rodríguez-Casal, A.: Set estimation under convexity-type assumptions. Ann. Inst. H. Poincaré Probab. Statist. **43**, 763–774 (2007)

[24] Rodríguez-Casal, A., Saavedra-Nieves, P.: A fully data-driven method for estimating density level sets. arXiv: 1411.7687v1 (2014)

[25] Tsybakov, A.B.: On nonparametric estimation of density level sets. Ann. Statist. **25**, 948–969 (1997)

[26] Walther, G.: On a generalization of Blaschke's rolling theorem and the smoothing of surfaces. Math. Meth. Appl. Sci. **22**, 301–316 (1999)

[27] Walther, G.: Granulometric smoothing. Ann. Statist. **25**, 2273–2299 (1997)

Chapter 8
Pseudo-metrics as Interesting Tool in Nonparametric Functional Regression

Laurent Delsol and Aldo Goia

Abstract The choice of the pseudo-metric in functional nonparametric regression is a crucial issue, since it has a direct impact on the kind of model one considers and on the efficiency of the estimation procedure. In this work a cross-validation approach to select the optimal pseudo-metric is illustrated and operationalized. Its practical performances are then evaluated by means of a Monte Carlo study.

8.1 Introduction

The so called functional data (such as, for instance, curves, surfaces, images) are nowadays commonly object of analysis both in applied and in theoretical Statistics. A lot has been done to deal with functional data since the pioneer works of [5], [6] and [9]: many multivariate methods and models have been extended to this field and new ideas have arisen and been developed (see, for instance, the recent special issues [2], [3], [7], [8] and the collective works [1] and [4]).

Looking at this literature, one can see how the regression models with a real-valued response and functional covariates have received a lot of attention, with a special attention to the nonparametric field which represents a fruitful domain of research (for an overview, see [6]).

In the nonparametric regression context, a central role is played by pseudo-metrics: since they are useful tools to capture the information contained in the explanatory curve, it is crucial to select the right one to obtain relevant results. This is usually done by cross-validation. However, only a partial theoretical result has been provided by [10] to prove its efficiency.

Laurent Delsol (✉)
Université d'Orléans, Rue de Chartres, B.P. 6759, FR-45067 Orléans cedex 2, France
,e-mail: Laurent.Delsol@univ-orleans.fr

Aldo Goia
Università del Piemonte Orientale, Via Perrone 18, 28100, Novara, Italia,
e-mail: aldo.goia@uniupo.it

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_8

The aim of the present work is to deepen this idea and extend it to a more general context. The goal is to explore its potential use to extract the informative features of the explanatory curve and hence construct more accurate estimates: after introducing the notation and the cross-validation idea (see Section 8.2), a practical criterion to select the optimal pseudo-metrics is illustrated (see Section 8.3). Finally, a simulation study illustrates the practical behaviour of the introduced approach in two special cases (see Section 8.4).

8.2 Pseudo-metrics use in Non-parametric Regression on Functional Variable

Consider the following regression model

$$Y = r(X) + \epsilon, \tag{8.1}$$

in which Y is a real-valued random variable, X is a random object lying in the space $\mathcal{F}$ of the real-valued functions defined on $[0, 1]$ which can be equipped with a family of pseudo-metrics $d \in \mathcal{D}$, r is a real-valued regression operator, and the residual ϵ satisfies $\mathbb{E}[\epsilon|X] = 0$.

Given a sample $(X_i, Y_i)_{1\leq i\leq n}$ drawn from (X, Y), in the nonparametric case, it is usual to make the kernel estimator of the regression operator r depend on a pseudo-metric d as follows (see e.g. [6]):

$$\widehat{r}_d(x) = \begin{cases} \sum\limits_{i=1}^{n} Y_i \dfrac{K\left(\dfrac{d(X_i,x)}{h_n}\right)}{\sum\limits_{i=1}^{n} K\left(\dfrac{d(X_i,x)}{h_n}\right)} & \text{if } \sum\limits_{i=1}^{n} K\left(\dfrac{d(X_i,x)}{h_n}\right) \neq 0 \\ \overline{Y} & \text{otherwise.} \end{cases} \tag{8.2}$$

The pseudo-metric d is usually understood as a simple tool to define a relevant estimator of the regression operator. However, its choice has a direct impact on the regression operator actually estimated. Indeed, the use of a given pseudo-metric d makes the estimator $\widehat{r}_d$ designed for the regression model induced on the quotient space $\mathcal{F}_d$:

$$Y = r_d(X) + \epsilon_d \tag{8.3}$$

for which $\mathbb{E}[\epsilon_d|\{Z, d(Z, X) = 0\}] = 0$ and hence $r_d(X) = \mathbb{E}[Y|\{Z, d(Z, X) = 0\}]$. When d is fixed and is not a metric, r_d has no reason to be equal to r and the estimator is usually not consistent.

The choice of the pseudo-metric is hence crucial: it has a direct impact on the kind of model one considers and the efficiency of the estimation procedure. A metric would ensure $r_d = r$ but may lead to low convergence rates (due to the curse of dimensionality). A pseudo-metric may be a relevant tool to extract the appropriate information from the explanatory curve and improve the convergence rate. However, its choice has to be done cleverly since a inappropriate pseudo-metric may lead to an inconsistent estimator (especially when $r_d \neq r$).

In practice, the pseudo-metric d is usually selected as a minimizer of the weighted cross-validation criterion:

$$\widehat{d} = \arg \min_{d_0 \in \mathcal{D}_n} \sum_{i=1}^{n} (Y_i - \widehat{r}_{d_0}^{(-i)}(X_i))^2 \pi(X_i), \tag{8.4}$$

where π is a positive weight function, $\widehat{r}_{d_0}^{(-i)}(X_i)$ is the leave-one out kernel estimator defined from the dataset $(X_{i'}, Y_{i'})_{1 \leq i' \leq n,\, i' \neq i}$ and $\mathcal{D}_n$ is the set of considered pseudo-metric which may depend on n.

Depending on the nature of the set $\mathcal{D}_n$, various specific situations can appear. Some examples are in the following:

1. **Unweighted pseudo-metric selection**

   a. **Informative points (or interval) selection**
   Given a non random discretization grid $(t_j)_{1 \leq j \leq p_n}$, consider the pseudo-metric family $d_{J_n}(X, x) = d((X(t_j), j \in J_n), (x(t_j), j \in J_n))$, where J_n is a subset of $\{1, \ldots, p_n\}$.

   b. **Informative support selection**
   One may also consider a continuous version of the previous one to detect the informative "support" J_n by considering the pseudo-metric family $d_{J_n}(X, x) = d(X_{|J_n}, x_{|J_n})$, where $X_{|J_n}$ (or $x_{|J_n}$) denotes the restriction of X (or x) to J_n which is an union of several non overlapped intervals of $[0, 1]$.

   c. **Informative basis selection**
   Given a non random functional basis $(\phi_j)_{1 \leq j \leq p_n}$ for $\mathcal{F}$, consider the pseudo-metric family $d_{J_n}(X, x) = d\big((< X, \phi_j >, j \in J_n), (< x, \phi_j >, j \in J_n)\big)$ where $J_n \subset \{1, \ldots, p_n\}$ is a set of selected basis functions.

2. **Weighted pseudo-metric selection**
   Consider a family of weighted pseudo-metrics $w \mapsto d_w$, with the weight w being a density.

   a. Take $p \in \mathbb{N}$, and consider the weighted $\mathbb{H}_w^{p,2}$ pseudo norms induced by the density functions w:

   $$d_w(X, x) = \sqrt{\int (X^{(m)}(t) - x^{(m)}(t))^2 w(t) dt}$$

   where $X^{(m)}$ denotes the m-th derivative of X.

   b. Consider the weighted $\mathcal{L}_w^2$ pseudo-norms induced by the discrete probability measure w:

$$d_w(X,x) = \begin{cases} \sqrt{\sum_{j=1}^{p_n}(X(t_j) - x(t_j))^2 w_j} & \text{for influent points selection;} \\ \sqrt{\sum_{j=1}^{p_n}(< X - x, \phi_j >)^2 w_j} & \text{for influent components selection.} \end{cases}$$

The nature of the selected density will be informative to understand which features of the explanatory curve X contain information on Y.

Note that the informative support selection (example 1.b above) is a particular case of continuous weighted pseudo-metric selection (example 2.a) in which the weight function w is selected among $\mathcal{U}(J_n)$ densities.

A theoretical study of the efficiency of the cross-validation selection rule, that is asymptotic equivalence of the MISE obtained with $\widehat{d}$ and the minimal one over all $\mathcal{D}_n$, is still in progress: it aims to extend some ideas illustrated in the work [10].

8.3 Optimal Pseudo-metric Selection in Practice

As explained in the previous section, the aim is to select the pseudo-metric leading to the smallest value of the cross-validation criterion (8.4). The algorithm to get or estimate this optimal pseudo-metric will differ from one situation to the other.

Consider first the selection of some points or components (examples 1.a and 1.c above). In these families, each pseudo-metric may be identified to a presence/absence vector $A_n \in \{0,1\}^{p_n}$. The optimization procedure may be done in an exhaustive way considering any configuration of this vector. When 2^{p_n} is huge, bottom-up or top-down procedures and discrete weighted pseudo-metrics (example 2.b above) may be relevant alternatives. Moreover, one may expect some sparsity of the vector A_n and hence add some LASSO type penalty to the cross-validation criterion (8.4).

Consider now the selection of the optimal weighted pseudo-metric (example 2) or the selection of the informative support (example 1.b). The main idea is to use a parametric approximation $\widetilde{w}$ of the weight function (except for discrete weighted pseudo-metrics, example 2.b, in which w is already a vector of real parameters). This parametrization may be done using a mixture of given densities $(f_j)_{1 \le j \le p_n}$ functions:

$$\widetilde{w} = \sum_{j=1}^{p_n} \theta_j f_j \qquad \text{with} \qquad \theta_j \ge 0, \ \sum_{j=1}^{p_n} \theta_j = 1 \tag{8.5}$$

The optimization problem is hence reduced to the selection of the optimum vector $(\theta_1, \dots, \theta_{p_n})$. This may be done using any relevant parametric optimization procedure.

Examples of densities that can be used to build the mixture 8.5 are the uniform ones, each one defined over p_n contiguous subintervals of $[0,1]$, or the normal ones truncated over $[0,1]$, having the means defined as p_n equispaced points in $[0,1]$ and the same standard deviation (which decreases as p_n increases).

8.4 Simulation Studies

This section presents the results on some experiments to show the performances of the proposed method: in particular the attention is focused on the informative support selection and the informative basis selection (see examples 1.b and 1.c above).

Each sample of curves X_i used in the simulations consists in $n = 200$ trajectories of a standard Brownian motion on $[0, 1]$, discretized over a grid of 1000 equispaced points $0 \leq t_1 < \cdots < t_{1000} \leq 1$. The errors ϵ_i are generated from independent centered Gaussian random variables $\mathcal{N}(0, \gamma)$ with γ equals to 0.2 times the standard deviation of $r(X)$ in order to control the signal-to-noise ratio.

Informative Support Selection

For each curve, define the mean values m_0 and m_1 over two intervals $(a_0, b_0] = (0.2, 0.3]$ and $(a_1, b_1] = (0.8, 0.9]$:

$$m_k(X_i) = \frac{1}{\#T_k} \sum_{j \in T_k} X_i(t_j), \qquad k = 0, 1,$$

where $T_k = \left\{ j : a_k < t_j \leq b_k \right\}$.

Consider $\phi(X_i) = 0.7 m_0(X_i) + 0.3 m_1(X_i)$ and define:

$$Y_i = \sin(\pi\phi(X_i)) + \epsilon_i \qquad i = 1, \ldots, 200.$$

The aim is to check if the proposed data-driven choice may lead to relevant identification of the informative parts of the curve. We search for the optimum weight θ_j in (8.5) with f_j uniform densities over $((j-1)/p_n, j/p_n]$, $j = 1, \ldots, p_n$:

$$\widetilde{w}(t, \theta) = \sum_{j=1}^{p_n} \theta_j \mathbb{I}_{\left(\frac{j-1}{p_n}, \frac{j}{p_n}\right]}(t)$$

with for all $1 \leq j \leq p_n$, $\theta_j \geq 0$ and $\sum_{j=1}^{p_n} \theta_j = 1$.

In Figure 8.1 one estimate of the weight function $\widetilde{w}$ with $p_n = 20$ is shown: one can see that the informative parts of the curves (highlighted in grey) are quite well identified.

In order to evaluate the accuracy of the cross-validation selection rule, the MISEs for 100 Monte Carlo samples are computed with $p_n = 10, 20, 30$. In Figure 8.2 these results are compared with the ones obtained when the classical (unweighted) $\mathcal{L}^2$ norm and the weighted $\mathcal{L}^2_{w_0}$ norm involving the true weight function w_0 are used. Looking at the boxplots, one can appreciate the good performances of the proposed approach: the gain with respect to the use of the $\mathcal{L}^2$ norm is considerable and appears not very sensitive to the choice of p_n.

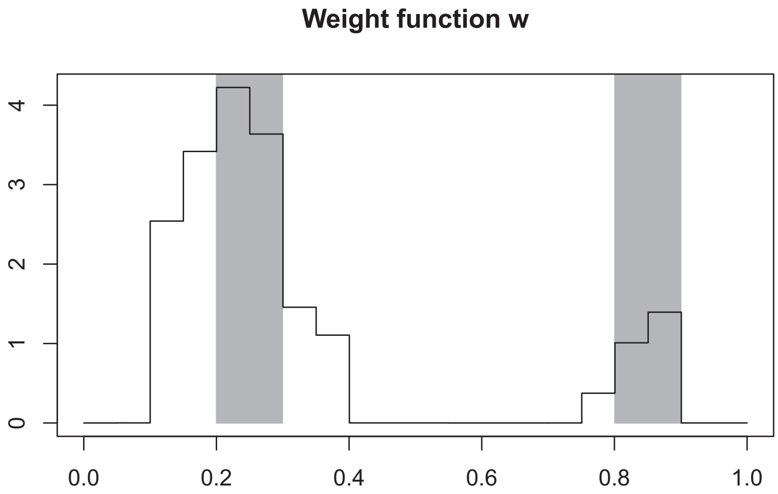


Fig. 8.1 An example of the estimated weight function w when $p_n = 20$

Informative Basis Selection

Consider the eigenvalues λ_j and the associated eigenfunctions ψ_j of the covariance operator of a Brownian motion:

$$\lambda_j = \frac{1}{\left(j - \frac{1}{2}\right)^2 \pi^2} \qquad j = 1, 2, \ldots$$

$$\psi_j(t) = \sqrt{2} \sin\left(\pi \left(j - \frac{1}{2}\right) t\right), \quad t \in [0, 1], \qquad j = 1, 2, \ldots$$

Define the regression model:

$$Y_i = [\sin(4\pi\varphi_2(X_i)) + \sin(4\pi\varphi_4(X_i))]^2 + \epsilon_i \qquad i = 1, \ldots, 200$$

where $\varphi_k(X_i) = \sqrt{\lambda_k}\,\langle\psi_k, X_i\rangle$, $k = 2, 4$, and estimate it nonparametrically using the pseudo-metric family introduced in the example 1.c with $p_n = 6$ (one deals with 64 possible configurations). Since in practice the basis in unknown, the eigenfunctions are estimated from the empirical covariance operator.

Table 8.1 collects the means and the standard deviations of the MISEs computed over 100 Monte Carlo replications, using some selections of the basis functions $(\psi_j)_{1 \le j \le 6}$. Only the best results, ordered decreasing according to the means, are

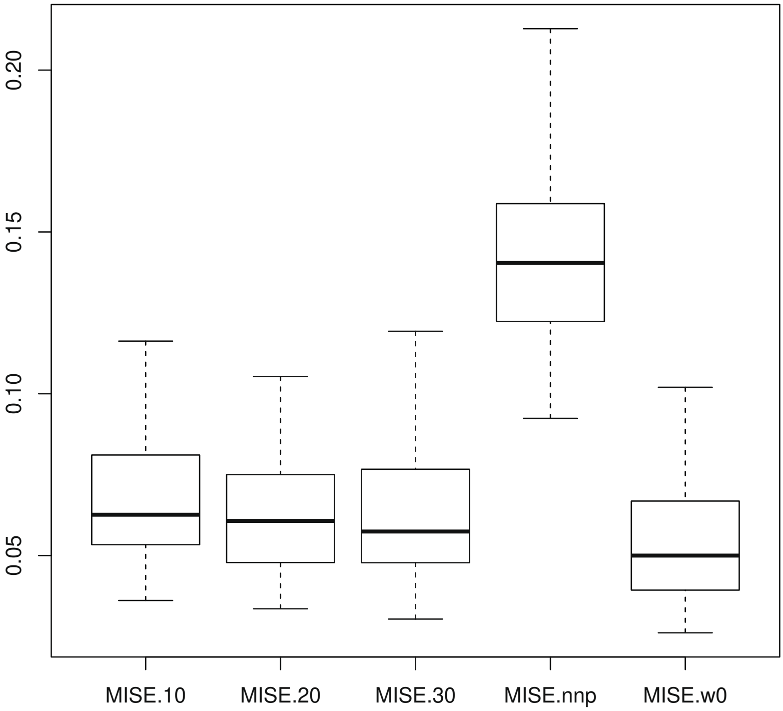


Fig. 8.2 The estimated MISEs when $p_n = 10, 20, 30$ (MISE.10, MISE.20 and MISE.30 respectively) and the ones when one uses the $\mathcal{L}^2$ and $\mathcal{L}^2_{w_0}$ norms (MISE.nnp and MISE.w0 respectively).

Table 8.1 Mean and standard deviation of MISEs over 100 MC simulations.

J_n	Mean	St.Dev.
$\mathbf{\{2,4\}}$	**1.589**	**0.293**
$\{2,4,6\}$	1.610	0.294
$\{2\}$	1.614	0.301
$\{2,4,5\}$	1.619	0.293
$\{2,5\}$	1.638	0.299
$\{1,2,3,4,5,6\}$	2.617	0.522

reported besides the one calculated when all the first 6 eigenfunctions are taken to evaluate the classical PCA-pseudo-metric. Reading the results, it emerges that the best performances are obtained when $J_n = \{2,4\}$, that is when only the informative basis elements φ_2 and φ_4 are exploited in the pseudo-metric.

Acknowledgements The authors want to thank an anonymous referee and the editors whose constructive comments and suggestions led to improvements in the manuscript. A. Goia thanks Università del Piemonte Orientale for the financial support and Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

References

[1] Aneiros, G., Bongiorno, E.G., Cao, R., Vieu, P.: Functional Statistics and Related Fields. Springer (2017)

[2] Aneiros, G., Cao, R., Fraiman, R., Vieu, P.: Editorial for the Special Issue on Functional Data Analysis and Related Topics. J. Multivar. Anal. **170**, 1–2 (2019)

[3] Aneiros, G., Cao, R., Vieu, P.: Editorial for the Special Issue on Functional Data Analysis and Related Topics. Comput. Stat. **34**, 447–450 (2019)

[4] Bongiorno, E.G., Goia, A., Salinelli, E., Vieu, P. (Eds.): Contributions in Infinite-Dimensional Statistics and Related Topics. Società Editrice Esculapio (2014)

[5] Bosq, D.: Linear Processes in Function Spaces. Theory and Applications. Lecture Notes in Statistics, Springer (2000)

[6] Ferraty, F. and Vieu, P.: Nonparametric Functional Data Analysis. Springer Series in Statistics. Springer, New York (2006)

[7] Goia, A., Vieu, P.: Editorial – An introduction to recent advances in high/infinite dimensional Statistics. J. Multivar. Anal. **146**, 1–6 (2016)

[8] Kokoszka, P., Oja, H., Park, B., Sangalli, L.: Special issue on functional data analysis. Econ. Stat. **1**, 99–100 (2017)

[9] Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, 2nd Edition. Springer Series in Statistics. Springer, New York (2005)

[10] Timmermans, C., Delsol, L., von Sachs, R.: Using Bagidis in nonparametric functional data analysis: Predicting from curves with sharp local features. J. Multivar. Anal. **115**, 421–444 (2013)

Chapter 9
Testing a Specification Form in Single Functional Index Model

Laurent Delsol and Aldo Goia

Abstract In this paper we propose a test of specification in functional regression with scalar response exploiting a semi-parametric approach. Once the test statistics is defined, its finite sample performances are analyzed through a simulation study.

9.1 Introduction

A wide part of the recent methodological and applied statistical literature is devoted to the study of functional regression models where relation between a random function X, defined on an interval, and a real response Y is described by:

$$Y = r[X] + \mathcal{E}$$

where r is a functional which models the conditional expectation of Y given X and $\mathcal{E}$ is a centered random error. This kind of model appears in many scientific domains and has been analyzed under different specifications of the regression operator r or in a full nonparametric context. For a review, see for instance the monographes [9], [14] or [19], or some works in collections of recent contributions ([1], [5]), or in special issues ([2], [3], [12], [15]).

In this framework, the check of the specification of the regression operator is a very important problem: the interest toward this issue is testified by a wide literature on structural testing procedures (see, for instance, [4], [6], [7], [8], [18]).

A useful help in this research field can come from the semi-parametric regression approaches. They combine flexibility and interpretability avoiding some dimensionality problems that can occour in the full nonparametric context. To have a partial

Laurent Delsol
Université d'Orléans, Rue de Chartres, B.P. 6759, FR-45067 Orléans cedex 2, France,
e-mail: Laurent.Delsol@univ-orleans.fr

Aldo Goia (✉)
Università del Piemonte Orientale, Via Perrone 18, 28100, Novara, Italy,
e-mail: aldo.goia@uniupo.it

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_9

idea on the richness of this research area, one can see the general presentation of semi-parametric approaches [13] and some recent papers and references therein (see, e.g. [10], [11], [17]).

In this work one exploits the Single Functional Index Model (SFIM) which postulates that, in the case of a predictor X valued in a Hilbert space $\mathcal{H}$ equipped with an inner product $\langle\cdot,\cdot\rangle$, the regressor operator can be written

$$r\left[X\right] = g(\langle X, \theta\rangle)$$

where g is a real unknown link function, and θ is an unknown element in $\mathcal{H}$ with unit-norm. So far, various techniques have been introduced to estimate g and θ (see e.g. [10]).

One of the main benefits of SFIM is that it makes possible to bring back an infinite dimensional problem to a one dimensional framework by projecting along the direction θ and then to graphically visualize an estimate of g, suggesting in this way the nature of the link between X and Y. For instance, a linear shape of the plot drives to a linear specification of the regression model. Hence, one can build a test procedure to check if a target specification g_0 for g, depending on some real parameter, is compatible with the observed dataset at a given significance level.

The aim of this work is to develop the test procedure: after illustrating the framework and the basic principle of the test (see Section 9.2), a test statistic based on a kernel approach is introduced (see Section 9.3). The performances of the test procedures on finite sample sizes are analyzed by means of a Monte Carlo simulation (see Section 9.4).

9.2 Notation and Basic Principle

Let X be a random variable (r.v. in the following) mapping in a Hilbert space $\mathcal{H}$ equipped with the inner product $\langle\cdot,\cdot\rangle$ and associated norm $\|\cdot\|$, and Y a real r.v. and state the following SFIM assumption

$$Y = g(\langle X, \theta\rangle) + \mathcal{E} \tag{9.1}$$

where $g : \mathbb{R} \longrightarrow \mathbb{R}$, $\theta \in \mathcal{H}$ with $\|\theta\|^2 = 1$ for identifiability, and $\mathcal{E}$ is a centered real r.v. with variance σ^2.

One wants to to test the following hypothesis:

$$H_0 : g\left(u\right) = g_0\left(\beta, u\right), \ \forall u \qquad \text{vs.} \qquad H_1 : \exists u : g\left(u\right) \neq g_0\left(\beta, u\right)$$

where $g_0\left(\beta, \cdot\right) : \mathbb{R} \longrightarrow \mathbb{R}$ is a known function, measurable w.r.t. the σ-algebra generated by X, depending on the parameter $\beta = \left(\beta_0, \beta_1, \ldots, \beta_p\right) \in \mathbb{R}^{p+1}$. To fix the ideas, the above setting includes the possibility of testing the linearity of the regression by specifying $g_0\left(\beta, u\right) = \beta_0 + \beta_1 u$.

Under the null hypothesis, for any positive weight function W, one has:

$$\mathbb{E}\left[\left(g(\langle X, \theta\rangle) - g_0\left(\beta, \langle X, \theta\rangle\right)\right)^2 W\left(X\right)\right] = 0 \tag{9.2}$$

Since, by the model (9.1), $g(\langle X, \theta \rangle) = \mathbb{E}[Y|X]$, then under H_0 one gets

$$g(\langle X, \theta \rangle) - g_0(\beta, \langle X, \theta \rangle) = \mathbb{E}[Y - g_0(\beta, \langle X, \theta \rangle)|X] = \mathbb{E}[\mathcal{E}|X]$$

and hence Equation (9.2) writes

$$\mathbb{E}\left[\mathbb{E}[\mathcal{E}|X]^2 W(X)\right] = \mathbb{E}[\mathcal{E}\mathbb{E}[\mathcal{E}|X]W(X)] = 0$$

whereas, under the alternative hypothesis one gets $\mathbb{E}[\mathcal{E}\mathbb{E}[\mathcal{E}|X]W(X)] > 0$.

As W can be chosen arbitrarily, one can take $W = f_\theta > 0$, the density distribution function of $\langle X, \theta \rangle$, a choice which helps later to simplify the test statistic expression. To implement a test procedure it is enough to introduce an empirical version of $\mathbb{E}[\mathcal{E}\mathbb{E}[\mathcal{E}|X] f_\theta(\langle X, \theta \rangle)]$ based on a sample drawn from (X, Y): one tends to reject H_0 if such quantity is significantly far from zero. The idea was introduced firstly in [21] for the multivariate setting.

9.3 The Test Statistic

Consider a sample $(X_i, Y_i), i = 1, \ldots, n$, of i.i.d. replications of (X, Y) and an estimator $\widehat{\theta}$ of θ. Let $\widehat{\beta}$ be some estimators of β, obtained under H_0 by regressing Y_i against the transformations of $g_0\left(\left\langle X_i, \widehat{\theta}\right\rangle\right)$ (in particular, if g_0 is affine, one regresses Y_i against $\left\langle X_i, \widehat{\theta}\right\rangle$). Finally define $\widehat{\mathcal{E}}_i = Y_i - g_0\left(\widehat{\beta}, \left\langle X_i, \widehat{\theta}\right\rangle\right)$.

At this stage, as said in Section 9.2, we have to provide a test statistic as the emprical version of $\mathbb{E}[\mathcal{E}\mathbb{E}[\mathcal{E}|X] f_\theta(\langle X, \theta \rangle)]$. To do this, we consider some non-parametric kernel estimates of $\mathbb{E}[\mathcal{E}|X]$ and f_θ at the point X_i and, by combining them, we get

$$\begin{aligned} T_n &= \frac{1}{n} \sum_{i=1}^{n} \widehat{\mathcal{E}}_i \left(\sum_{\substack{j=1 \\ j \neq i}}^{n} \widehat{\mathcal{E}}_j \frac{K_{ij}}{\sum_{j \neq i} K_{ij}} \right) \frac{1}{(n-1)h} \sum_{j \neq i} K_{ij} \\ &= \frac{1}{n(n-1)h} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \widehat{\mathcal{E}}_i \widehat{\mathcal{E}}_j K_{ij} \end{aligned}$$

where $K_{ij} = K\left(\left(d\left(X_i, X_j\right)\right)/h\right)$ and K is a kernel function, h is a suitable bandwith and d a suitable semi-metric. Recalling that we are working in the frame of SFIM, it is natural to choose:

$$d\left(X_i, X_j\right) = \left|\langle \theta, X_i \rangle - \left\langle \theta, X_j \right\rangle\right| = \left|\left\langle \theta, X_i - X_j \right\rangle\right|$$

that is, the semi-norm based on projection over the (one-dimensional) subspace spanned by θ, the first projection pursuit direction. Since θ is unknown, it is replaced by $\widehat{\theta}$, providing the test statistic:

$$T_n = \frac{1}{n\left(n-1\right)h} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \widehat{\mathcal{E}}_i \widehat{\mathcal{E}}_j K_{ij}^{\widehat{\theta}}$$

where $K_{ij}^{\widehat{\theta}} = K\left(\left|\left\langle \widehat{\theta}, X_i - X_j \right\rangle\right| / h\right)$.

Following similar arguments as in [16] or [21] one can derive the following estimate for the variance of T_n:

$$S_n^2 = \frac{2}{n\left(n-1\right)h} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \widehat{\mathcal{E}}_i^2 \widehat{\mathcal{E}}_j^2 \left(K_{ij}^{\widehat{\theta}}\right)^2 .$$

This allows to obtain the studentized test statistic version $n\sqrt{h}T_n/S_n$ for which a theoretical study is still in progress.

9.4 Finite Dimensional Performances of the Test

In this section the performances of the test by using Monte Carlo experiments under different experimental conditions are analyzed: the empirical power is computed as the frequency of times in which the test reject the null hupothesis over the number of replications. The critical region of the test at the level α is based on the Gaussian approximation of the null distribution: one rejects the null hypothesis whenever the value of the studentized test statistics is greather than the quantile of order $1-\alpha$ of the standard normal distribution. That approximation is based on heuristic arguments deriving on evidences from simulations and could be supported theoretically by results in [18].

The data used in the simulations are generated according to the following SFIM model:

$$Y_i = g\left(\langle \theta, X_i \rangle\right) + \sigma \mathcal{E}_i \qquad i = 1, \ldots, n$$

where:

- $\mathcal{E}_i$ are i.i.d. standard normal r.v.s and $\sigma^2 = \rho^2 Var\left(g\left(\langle \theta, X \rangle\right)\right)$ with ρ controlling the signal-to-noise ratio (we use $\rho^2 = 0.2$ and the variance is estimated for each sample);
- The functional covariate obeys to:

  $$X_i\left(t\right) = A_i + B_i t^2 + C_i \exp\left(t\right) + \sin\left(D_i 2\pi t\right) \qquad t \in \left[-1, 1\right]$$

  where A_i, D_i, C_i and D_i are r.v.s independent and uniformly distributed over $(-1, 1)$, so that the random process is centered; each trajectory is discretized over a grid of 100 equispaced design points;
- The functional coefficient is:

  $$\theta\left(t\right) = \kappa \cos(2\pi t^2) \qquad t \in \left[-1, 1\right]$$

being κ the normalization constant such that $||\theta|| = 1$; in such a way, the r.v. $\langle \theta, X \rangle$ is centered and bounded.
- The sample sizes used are: $n = 50, 100, 200$;
- All the integrals are approximated by summations.

In the following we present the results when one tests linearity and a nonliner case at the nominal level $\alpha = 0.05$. For what concerns the estimation of the SFIM we use the first step of the approach proposed in [10], which combines a spline approximation of the functional coefficient θ and the one-dimensional Nadaraya-Watson approach: the bandwidth in this latter is selected by a leave-one-out CV approach, whereas the approximation of θ is based on cubic splines with 5 knots. The bandwidth h used in computing the test statistics is selected by the unbiased cross-validation approach when one estimates the density f_θ.

9.4.1 Testing Linearity

Consider the following null hypothesis:

$$H_0 : g(u) = u$$

The power of the test is evaluated from 1000 Monte Carlo replications for the following alternatives:

$$H_1^{(1)}(\gamma) : g(u) = u + \gamma u^2$$
$$H_1^{(2)}(\gamma) : g(u) = u + \gamma \sin(3\pi u)/10$$

with $\gamma = 0.5, 0.75, 1$ (when $\gamma = 0$ the empirical level is computed). The results of the experiments are collected in Table 9.1. Despite one uses an asymptotic approximation for the null distribution, the empirical level is rather close to the theoretical one, also for a relatively small sample size. As expected, the further one moves away from the linearity by increasing γ, the greater is the estimated power, both in the quadratic and the sinusoidal case. Anyway, very good are the performances when $n = 200$, also with a relatively modest departure from the null hypothesis.

9.4.2 Testing a Cubic Link

In this second experiment, one considers the null hypothesis of a cubic link:

$$H_0 : g(u) = u^3$$

The empirical level and the empirical power have been computed from 1000 MC replications for the following family of alternatives:

$$H_1(\delta) : g(u) = u^3 + \delta u \qquad \delta = 0.02, 0.04, 0.06, 0.08$$

Table 9.1 Estimated level (first line) and power for the test of linearity under different experimental conditions.

	Sample size		
Alternatives	50	100	200
$H_1^{(\star)}$ (0)	0.057	0.053	0.054
$H_1^{(1)}$ (0.5)	0.143	0.236	0.434
$H_1^{(1)}$ (0.75)	0.276	0.507	0.828
$H_1^{(1)}$ (1)	0.439	0.811	0.984
$H_1^{(2)}$ (0.5)	0.209	0.397	0.750
$H_1^{(2)}$ (0.75)	0.381	0.710	0.967
$H_1^{(2)}$ (1)	0.598	0.911	1.000

Table 9.2 Estimated level (first line) and power for testing cubic specification under different experimental conditions.

	Sample size		
Alternatives	50	100	200
H_1 (0)	0.069	0.061	0.056
H_1 (0.02)	0.186	0.263	0.419
H_1 (0.04)	0.394	0.624	0.891
H_1 (0.06)	0.587	0.846	0.993
H_1 (0.08)	0.702	0.957	1.000

(for $\delta = 0$ the empirical level is given). The results of the experiments are collected in Table 9.2.

Also in this second experiment, the obtained results are rather good. The estimated level is slightly higher than the nominal one, providing a liberal test in particular for relatively small sample size. About the estimated power, it increases with the sample size n and with δ, following a coherent behaviour with the departure from the null hypothesis. In conclusion, one has an empirical evidence that the Gaussian approximation of the null distribution works reasonably well.

9.5 Concludings

In this paper a test procedure for checking the validity of SFIM is introduced and its performances under some special alternatives are analyzed for finite dimensional samples: the obtained results, based on a Gaussian approximation of the null distribution, appear promising.

The work opens towards at least two future lines of research. The first one is to provide a rigorous theoretical study: in particular, to get the asymptotic null distribution and the power of the test, at least under some alternatives, by exploiting and adapting ideas and mathematical tools proposed in [18]. The second line is to improve the performances of the test by using a threshold value obtained with a bootstrap procedure. In practice, once bootstrap samples are generated (under the null hypothesis) and for each of them the test statistic has been computed, the empirical distribution of the latter is available. Hence its quantile of order $1 - \alpha$ can be evaluated and used in defining the critical region of the test.

Acknowledgements The authors wish to thank an anonymous referee for his helpful comments and suggestions. A. Goia thanks Università del Piemonte Orientale for the financial support and Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

References

[1] Aneiros, G., Bongiorno, E.G., Cao, R., Vieu, P.: Functional Statistics and Related Fields. Springer (2017)

[2] Aneiros, G., Cao, R., Fraiman, R.,Vieu, P.: Editorial for the Special Issue on Functional Data Analysis and Related Topics. J. Multivar. Anal. **170**, 1–2 (2019).

[3] Aneiros, G., Cao, R., Vieu, P.: Editorial on the special issue on Functional Data Analysis and Related Topics. Comput. Stat. **34**(2), 447–450 (2019)

[4] Aneiros, G., Vieu, P.: Testing linearity in semi-parametric functional data analysis. Comput. Stat. **28**(2), 413–434 (2013)

[5] Bongiorno, E.G., Goia, A., Salinelli, E., Vieu, P. (Eds.): Contributions in Infinite-Dimensional Statistics and Related Topics. Società Editrice Esculapio (2014)

[6] Bücher, A., Dette, H., and Wieczorek, G.: Testing model assumptions in functional regression models. J. Multivar. Anal. **102**, 1472–1488 (2011)

[7] Cuesta-Albertos, J.A., García-Portugués, E., Febrero-Bande, M., González-Manteiga, W.: Goodness-of-fit tests for the functional linear model based on randomly projected empirical processes. Ann. Stat. **47**(1), 439–467 (2019)

[8] Delsol, L., Ferraty, F., Vieu, P.: Structural test in regression on functional variables. J. Multivar. Anal. **102**(3), 422-447 (2011)

[9] Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis. Springer Series in Statistics. Springer, New York (2006)

[10] Ferraty, F., Goia, A., Salinelli, E., Vieu, P.: Functional Projection Pursuit Regression. Test **22**, 293–320 (2013)

[11] Goia, A., Vieu, P.: A partitioned Single Functional Index Model. Comput. Stat. **30**, 673–692 (2015)

[12] Goia, A., Vieu, P.: Editorial – An introduction to recent advances in high/infinite dimensional Statistics. J. Multivar. Anal. **146**, 1–6 (2016)

[13] Härdle, W., Müller, N., Sperlich, S., Werwatz, A.: Nonparametric and semiparametric models. Springer (2004)

[14] Horvath, L., Kokoszka, P.: Inference for Functional Data with Applications. Springer Series in Statistics. Springer, New York (2012)

[15] Kokoszka, P., Oja, H., Park, B., Sangalli, L.: Special issue on functional data analysis. Econ. Stat. **1**, 99–100 (2017)

[16] Li, Q., Wang, S.: A simple consistent bootstrap test for a parametric regression function. J. Econom. **87**, 145–165 (1998)

[17] Novo, S., Aneiros-Pérez, G., Vieu, P.: Automatic and location-adaptive estimation in functional single-index regression. J. Nonparametr. Stat. **31**, 1–29 (2019)

[18] Patilea,V., Sánchez-Sellero, C., Saumard, M.: Testing the predictor effect on a functional response. J. Am. Stat. Assoc. **111**, 1684–1695 (2016)

[19] Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, 2nd Edition. Springer Series in Statistics. Springer, New York (2005)

[20] Silverman, B.W.: Density Estimation. London, Chapman and Hall (1986)

[21] Zheng, X.: A consistent test of functional form via nonparametric estimation technique. J. Econom. **75**, 263–289 (1996)

Chapter 10
A New Method for Ordering Functional Data and its Application to Diagnostic Test

Graciela Estévez-Pérez and Philippe Vieu

Abstract This contribution proposes a new ordering method for functional data which could be a starting point for developing new advances in problems for which the ordering of curves is of interest. This method is used to construct a diagnostic test, based on the functional version of ROC curves, for situations when the observed biomarker is a functional variable.

10.1 Introduction

A growing number of fields are handling functional data, that is, data coming from realizations of random elements taking values in infinite-dimensional spaces (i.e. in functional spaces). In the last decades, multiple statistical procedures have been extended to functional data context emerging a new area in Statistic known as functional data analysis (FDA). For an overview of this topic see, for instance, [19], [9], [12] and the recent survey discussions by [4], [11] or [1]. However, any functional procedure which needs to estimate a functional distribution function is no direct, because it requires establishing some ordering method for elements of infinite-dimensional spaces. This is one of the challenge that this work aims to front.

To be exact, as part of the main purpose of constructing a diagnostic test based on functional biomarkers, we are interested in achieving an idea of when a functional data precedes to another in some sense. For that, after some quick discussions about ranking of data (Sect. 10.2) and application of ROC curves to diagnostic test (Sect. 10.3), we are presenting our general methodology for ordering functional data in Sect. 10.4. Some simple examples are also developed in this section to show how

Graciela Estévez-Pérez (✉)
Departamento de Matemáticas, Universidade da Coruña, Spain, e-mail: graci@udc.es

Philippe Vieu
Institut de Mathématiques, Université Paul Sabatier, Toulouse, France,
e-mail: philippe.vieu@math.univ-toulouse.fr

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_10

the procedure works in practice. Finally, Sect. 10.5 briefly outlines our main purpose of constructing a diagnostic test based on a functional biomarker.

10.2 Some Discussion about Ranking: From One to Infinite Dimensions

It is well known that several unambiguous orderings can be consider in one-dimensional spaces. However, for multivariate data the first attempt of ordering consisted in projecting the multidimensional space in a suitable one dimensional subspace, so it retains as much information as possible of data. Several approaches for choosing this subspace have been proposed from [2], arising different sub-ordering methods so-called *reduced ordering* (or *R-ordering*). See, for example, [8], [15] and [7].

Nowadays, one of the most popular ordering approach in multidimensional spaces is based on the notion of data depth. A multivariate depth is a function that provides a P-based center-outward ordering of points in $\mathbb{R}^d$, where P is a probability distribution on $\mathbb{R}^d$ (see [20]), so that high values of depth indicate points that are central relative to the probability distribution P and lower depths indicate peripheral points for P. Therefore, one can compute the depths of all the sample points and order them according to decreasing depth values, obtaining a ranking of the sample points from the center outward. For an overview on multivariate depths see, for example, [22].

In infinite-dimensional spaces, several notions of functional depth have been recently defined ([10], [5], [3], [6] and [21]). However, this partial ordering, which induces a centre-outward ranking of the observations, does not distinguish the location relative to the data center for two observations which have the same depth. Thus, as our goal is not to order observations with respect to a center (the deepest point), but to identify whether a curve precedes to another in some sense, an alternative ordering methodology is needed for functional data. This is what we describe in Sect. 10.4.

10.3 Some Discussion about ROC Curves and Diagnostic Test

The receiver operating characteristic (ROC) curve is the most popular tool to evaluate the accuracy of diagnostic tests when continuous real variables are used as markers, thus it is of great importance in clinical practice and medical research. There are numerous articles in the literature on parametric and nonparametric statistical methods for estimating ROC curves ([18] and [17] among others) and their most common summary measure, the area under the curve (AUC) (see for example, [16]).

The technological advances of the last decades are allowing to observe, in some settings, markers with more complex structures. Thus, the need to extend the ROC methodology to functional context becomes each day more important. The FDA community has started debating recently on the development of diagnostic analysis and curves ROC estimation techniques but, at our knowledge, only when functional data are involved as covariables ([13] and [14]). The extension of multivariate ideas to this situation is more easy since the conditional distribution function is still an un-functional one. Note that any diagnostic procedure based on ROC curves needs to estimate the distribution function of the considered random biomarker. The reason for this lack of development for ROC analysis with functional sample is probably linked with the difficulty of establishing a suitable ordering in infinite-dimensional spaces. The procedure for ordering infinite-dimensional data, which will be proposed below, will enable to create a diagnostic test based on a functional biomarker, as we comment in Sect. 10.5.

10.4 A Flexible Functional Ordering Method

Following ideas close to R-ordering (see [7] and its references), we will propose next a pre-ordering method for functional data which allows identifying whether a curve precedes to another in some sense.

The problem of ordering functional data can be expressed as follows. Given $\chi_1(t) \equiv \chi_1, \chi_2(t) \equiv \chi_2; t \in T$ two random points of a functional space (E, d), where d is a semi-metric and $T = [0, 1]$ without loss of generality, we are interested in defining an arrangement of the form $\chi_1 \preceq \chi_2$, where the symbol "$\preceq$" means "*precede to*". We propose an ordering procedure which involves the preordering of the functional space E with respect to an ordered curves set. Precisely, given χ_1 and χ_2 two random points of the functional space (E, d), the construction of the ordering methodology can be summarized by the three following steps:

First step Fix some subset $(E_1, \preceq_1)$ in E, in such a way that one can construct $\preceq_1$ to be a total order relation on E_1;

Second step Define $\eta(\cdot)$ to be a projection function of E into E_1;

Third step Say that χ_1 precedes to χ_2 in E (and write it as $\chi_1 \preceq \chi_2$), if and only if $\eta(\chi_1) \preceq_1 \eta(\chi_2)$ in E_1.

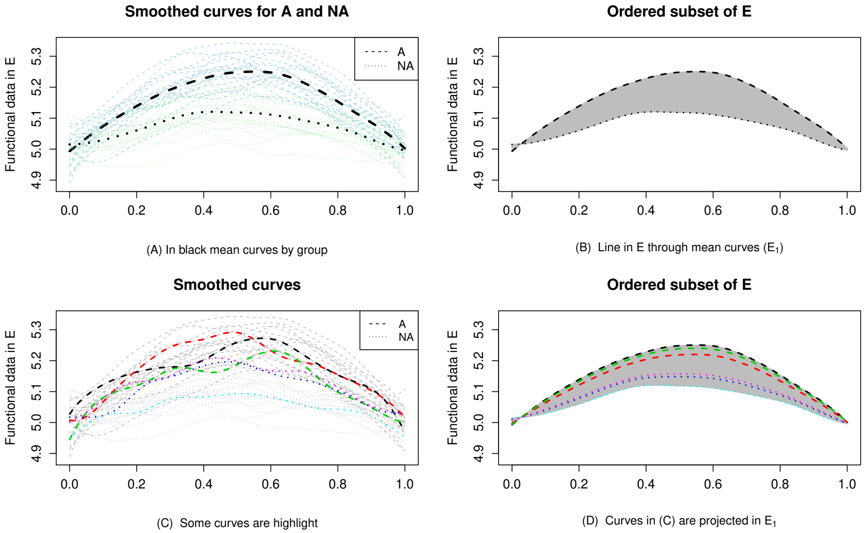
Note that $\preceq$ is a preorder relation in E since it satisfies the properties of reflexivity and transitivity but it is not an order since the antisymmetry is not verified.

This methodology is very flexible because many different choices can be made, both for the pilot ordered subset $(E_1, \preceq_1)$ as for the projector $\eta(\cdot)$. Several options are possible depending on the kind of statistical problem one has to deal with.

10.4.1 How Does the Classification Procedure Work in Practice?

One simulated example will be developed below to show how the ordering procedure works in practice. We address a supervised classification problem because the diagnostic test are the main focus of this paper, but other issues could be discussed.

Example: Ranking a grouped sample. In this example we show the ranking procedure's usefulness in a supervised classification problem of curves. For that, a set of $n_1+n_2 = 20+20 = 40$ curves were generated from two continuous time process with mean functions $\mu_1(t) = 5+t(1-t)$ and $\mu_2(t) = 5+t^{3/2}(1-t)^{3/2}$, for $t \in T = [0,1]$ ($n_1 = 20$ from A group (dashed line) and $n_2 = 20$ from NA group (dotted line)). As it can be seem in panel (A) of Figure 10.1, the curves were smoothed and represented by group membership together with the mean curves (thick line). E_1 has been chosen as the "line in E" through the sample means functions by group, that is, $E_1 = \{\chi_c(t) = c\overline{\chi_A}(t) + (1-c)\overline{\chi_{NA}}(t); t \in T, c \in [0,1]\}$ (see panel (B)). $\eta(\cdot)$ has been chosen as the function which projects E in E_1 searching for each $\chi \in E$ the closest point in E_1 in accordance with the semi-metric based on mplsr ([9]). To illustrate the behaviour of the ordering procedure three curves by group, randomly selected, were highlight in panel (C) and projected on E_1 (see panel (D)).



(A) In black mean curves by group

(B) Line in E through mean curves (E_1)

(C) Some curves are highlight

(D) Curves in (C) are projected in E_1

Fig. 10.1 Example of curves in E and their projections in E_1

By comparing the panels (C) and (D) in Fig. 10.1, it can be appreciated how the ordering procedure works and the distance between the projected curves. It is worth noting the discrimination in E_1 between curves coming from different population.

10.5 A Functional Diagnostic Test

The general functional ranking strategy, depicted before in Sect. 10.4, could be a starting point for developing new advances in any problem for which the ordering of curves is of interest. In fact, we have used it to construct a diagnostic test based on functional biomarkers. The main steeps of method are outlined below and more detailed information will be provide in the work presented during IWFOS 2020.

1. **The framework** Let $\chi \equiv \chi(.)$ be a functional biomarker and $\chi_1 = \chi/A$, $\chi_2 = \chi/NA$ be the conditioned functional variables by affected subjects (A) and non-affected ones (NA), which are valued in a semi-metric space (E, d). In addition, we suppose that the probability measures of random variables χ_1 and χ_2 differ by a shift $\Delta \in E$ in the location.

   We are in front of a discrimination problem and the data consist in n_1 and n_2 i.i.d. random samples as χ_1 and χ_2

   $$\{\chi_{ij}(t); i = 1, 2; j = 1, ..., n_i; t \in T\}$$

   where, without loss of generality, $T = [0, 1]$ and $n = n_1 + n_2$ denotes the total sample size.

2. **The classification rule** Our diagnostic method consists in finding a cutoff function or discrimination function $\chi_{c_0} \equiv \chi_{c_0}(t) \in E$, with $t \in T$, which separates Affected (A) and Non-Affected (NA) subjects properly.

   Step 1 We select $M_A \equiv M_A(\cdot)$ and $M_{NA} \equiv M_{NA}(\cdot)$ two curves of the functional space E as representatives of each group A and NA, respectively. Then, we consider the totally ordered set $(E1, \preceq_1)$, where E_1 is the line in E through functions M_A and M_{NA}, that is, the family of curves:

   $$E_1 = \{\chi_c(t) = cM_A(t) + (1 - c)M_{NA}(t); t \in T, c \in [0, 1]\}$$

   and $\preceq_1$ is the total order in E_1 induced by the natural ordering in $\mathbb{R}$.

   Step 2 We take the pilot ordered subset E_1 with the total order relation $\preceq_1$ and we consider the projection function $\eta : E \longrightarrow E_1$ that assigns to each curve in E the nearest element in E_1 when a semi-metric $d(,)$ is used. That is, the projection function is defined by:

   $$\eta(x) = \min\{y \in E_1/d(x, y) = \min\{d(x, y_1); y_1 \in E_1\}\}$$

   Note that $\eta(\cdot)$ allows to establish a preorder relation in E $(\preceq)$: we say that x_1 precedes to x_2 in E (and write it as $x_1 \preceq x_2$), if and only if $\eta(x_1) \preceq_1 \eta(x_2)$ in E_1.

   Step 3 Predetermined a cutoff curve ζ and assuming that curves with higher values are linked with Affected subjects, **an individual associated with a new functional data χ_0 is diagnosed as Affected (A) if** $\chi_0 \npreceq \zeta$. In this case we say that the test result is +, and we denote it by T^+.

As the classification result depends on the prefixed cutoff curve $\zeta \in E$, we propose to move continuously ζ in $\{\chi_c \in E_1; c \in [0,1]\}$ and to take the value $c_0 \in [0,1]$ such as the curve χ_{c_0} lets to discriminate A and NA in the most efficient way.

3. **How to select the optimal cutoff curve χ_{c_0}?** The accuracy of the test is defined by the pairs Specificity $(1-\alpha_c)$ and Sensitivity $(1-\beta_c)$ for each cutoff curve χ_c. If the value c is moved in $[0,1]$, the sensitivity and the specificity vary inversely, then to selecting the "best" cutoff curve χ_{c_0} it is needed a balance between these values. Thus, the functional version of the Receiver Operating Characteristic (ROC) curve is defined by the graph in the unit square representing for each $c \in [0,1]$ the pair $(\alpha_c, 1-\beta_c)$, which represents a global measure of the discriminatory ability of our functional diagnostic test along all cutoff curves χ_c.

   Based on data $\{\chi_{ij}(t); i=1,2; j=1,...,n_i; t \in T\}$, the most natural procedure for estimating the ROC curve is by its empirical version.

4. **An optimal data-driven procedure** For selecting **the optimal cutoff curve χ_{c_0} or discrimination curve**, we propose to take χ_{c_0} such that the pair of values $(1-\alpha_{c_0}, 1-\beta_{c_0})$ is as close as possible to the pair of perfect classification given by $(1,1)$. This is equivalent to searching **the point on the ROC curve closest to the** $(0,1)$ **point**, and this method is usually called the ***North-West corner* or the *Closest-to-(0,1) criterion***.

   Thus, individuals associated with a functional data $\chi \npreceq \chi_{c_0}$ are classified as Affected (positive test), whereas individuals with $\chi \preceq \chi_{c_0}$ are classified as Non-Affected (negative test).

From a comprehensive simulation study, we have observed that functional diagnostic test proposed is a good option to discriminate two groups and classify new subjects when the biomarker is functional. Several simulation scenarios whose mean functions present different degrees of separation and hence, different levels of difficulty to distinguish Affected subjects from Non-Affected ones, have been considered. We have also considered diverse sample sizes and several levels of variability intra and inter curves, to cover a wider range of real situations.

Acknowledgements This research/work has been supported by MINECO grant MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the ERDF.

References

[1] Aneiros, G., Cao, R., Fraiman, R., Genest, C., Vieu, P.: Recent advances in functional data analysis and high-dimensional statistics. J. Multivariate Anal.

170, 3–6 (2019)

[2] Barnett, V.: The ordering of multivariate data. Journal of the Royal Statistical Society, A **139**, 318–354 (1976)

[3] Cuesta-Albertos, J.A., Nieto-Reyes, A.: The random tukey depth. Computational Statistics and Data Analysis **52**, 4979–4988 (2008)

[4] Cuevas, A.: A partial overview of the theory of statistics with functional data. J. Statist. Plann. Inference **147**, 1–23 (2014)

[5] Cuevas, A., Febrero, M., Fraiman, R.: On the use of the bootstrap for estimating functions with functional data. Computational Statistics and Data Analysis **51**, 1063–1074 (2006)

[6] Cuevas, A., Fraiman, R.: On depth measures and dual statistics. A methodology for dealing with general data. Journal of multivariate analysis **100**, 753–766 (2009)

[7] D'Esposito, M.R., Ragozini, G.: A new R-ordering procedure to rank multivariate performances. Quaderni di Statistica **10**, 5–21 (2008)

[8] Eddy, W.F.: Ordering of multivariate data in computer Science and Statistics: the Interface. L. Billard Ed., Amsterdam, North-Holland, 25–30 (1985)

[9] Ferraty, F., Vieu, P.: Nonparametric functional data analysis: theory and practice. Springer Series in Statistics, Springer, New York (2006)

[10] Fraiman, R., Muniz, G.: Trimmed means for functional data. Test **10**, 419–440 (2001)

[11] Goia, A., Vieu, P.: An introduction to recent advances in high/infinite-dimensional statistics. J. Multivariate Anal. **146**, 1–6 (2016)

[12] Horváth, L., Kokoszka, P.: Inference for functional data with applications. Springer Series in Statistics, Springer, New York (2012)

[13] Inácio, V., González-Manteiga, W., Febrero-Bande, M., Gude, F., Alonzo, T.A.: Extending induced ROC methodology to the functional context. Biostatistics **13**(4), 594–608 (2012)

[14] Inácio, V., de Carvalho, M., Alonzo, T.A., González-Manteiga, W.: Functional Partial Area under the Curve Regression with an Application to Metabolic Syndrome Diagnosis. Annals of Applied Statistics **10**, 1472–1495 (2016)

[15] Korhonen, P., Siljamaki, A,: Ordinal principal component analysis. Theory and an application, Computational Statistics and Data Analysis, **26**, 411–424 (1998)

[16] Ma, H., Bandos, A.I., Rockette, H.E., Gur, D.: On use of partial area under the ROC curve for evaluation of diagnostic performance. Statistics in medicine **32**, 344–3458 (2013)

[17] Peng, L., Zhou, X.H.: Local linear smoothing of receiver operating characteristic (ROC) curves. Journal of Statistical Planning and Inference **118**, 129–143 (2004)

[18] Pepe, M.S.: The statistical evaluation of medical tests for classification and prediction. 1st ed. Oxford University Press, USA (2004)

[19] Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, 2nd edn. Springer Series in Statistics, Springer, New York (2005)

[20] Serfling, R.: Depth functions in nonparametric multivariate inference. In: Liu, R., Serfling, R., Souvaine DL (eds) Data depth: robust multivariate analysis, computational geometry and applications. DIMACS Series. American Mathematical Society, Providence, 1–16 (2006)

[21] Sguera, C., Galeano, P., Lillo, R.: Spatial depth-based classification for functional data. Test **23**, 725–750 (2014)

[22] Zuo, Y., Serfling, R.: General notions of statistical depth function. The Annals of Statistics, 461–482 (2000)

Chapter 11
A Functional Data Analysis Approach to the Estimation of Densities over Complex Regions

Federico Ferraccioli, Laura M. Sangalli, Eleonora Arnone and Livio Finos

Abstract In this work we propose a nonparametric method for density estimation over two-dimensional domains. Following a functional data analysis approach, we consider a penalized likelihood estimator, with a roughness penalty based on a differential operator. This approach allows for the estimation of densities on any planar domain, including those with complex boundaries or interior holes. We develop an estimation procedure based on finite elements. Thanks to the use of this numerical technique, the proposed method has great flexibility and high computational efficiency.

11.1 Introduction

It the recent years there has been an increasing cross-contamination of techniques from functional data analysis and from spatial data analysis; see, e.g., the special issue [19] and the review in [9]. Here in particular we consider the problem of estimating a density function f on a two-dimensional domain with a complex shape. For example, Figure 11.1 illustrates crime locations in the municipality of Portland, Oregon. The interest is to study the distribution of crimes in order to identify critical and dangerous areas in the city. Here the complex geographical conformation of the

Federico Ferraccioli (✉)
Dipartimento di Scienze Statistiche, Via Cesare Battisti, 241, 35121 Padova, Italy,
e-mail: ferraccioli@stat.unipd.it

Laura M. Sangalli
MOX-Dipartimento di Matematica, Piazza L. da Vinci, 32, 20133 Milano, Italy,
e-mail: laura.sangalli@polimi.it

Eleonora Arnone
MOX-Dipartimento di Matematica, Piazza L. da Vinci, 32, 20133 Milano, Italy,
e-mail: eleonora.arnone@polimi.it

Livio Finos
Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Via Venezia, 8, 35131 Padova, Italy, e-mail: livio.finos@unipd.it

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_11

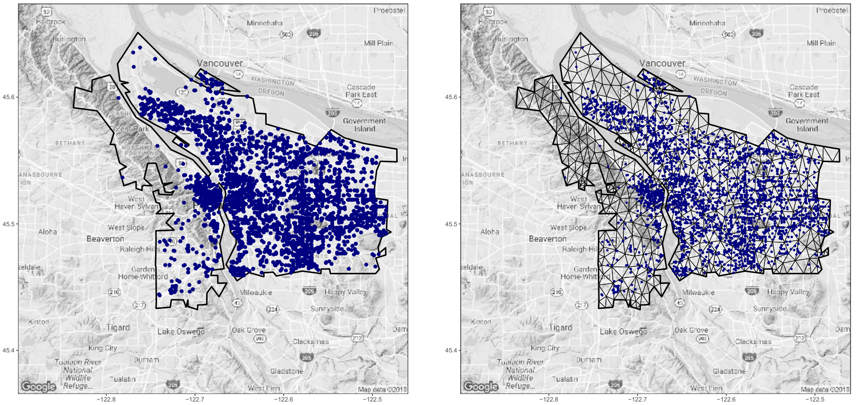


Fig. 11.1 On the left, points distributed over a complex domain with boundaries. On the right, the constrained Delaunay triangulation of the same domain.

domain, and in particular the presence of the river, is crucial in the study of the phenomenon. There is a clear difference between the West-side, characterized by a lower number of crimes, and the East-side, characterized by a higher number of crimes. The difference is particularly pronounced in the Northern and in the Southern part of the city.

The analysis of data observed over domains with complex shapes has lately drawn lots of attention. [21] and [26] propose smoothing methods with a regularization based on a differential operator; [22] extends these models to spatial regression and [2] considers two regularizing terms involving general partial differential equations; [3] deals with spatially dependent functional data over complex domains, and [18] tackles the case of object data. The problem of density estimation in this complex setting as not been addressed yet.

In this work we present a flexible density estimation method for data distributed over complex regions. The model formulation is based on a nonparametric likelihood approach, with a regularization that involves partial differential operators. In the univariate case, a similar approach is considered in [13] and later developed in [23]. In the multivariate setting, the proposal in [14] develops a spline model that can be used for simple tensorized domains. All these methods are nonetheless not easily generalizable to the case of complex multivariate domains.

The method we propose leverages on advanced numerical analysis techniques to address the estimation problem. In particular, we use the finite element method, often used in engineering applications to solve partial differential equations. The main advantage of these techniques consists in the possibility of considering spatial domains with complex shapes, instead of simple tensorized domains, as considered by [14] and by the other available methods for density estimation. Moreover, the proposed method for density estimation does not impose any shape constraint; on the contrary, it permits the estimation of fairly complex structures. In particular, thanks to the finite element formulation, the method is able to capture highly localized features, and lower dimensional structures such as ridges. This ability also makes the method particularly well suited in research areas such as density based clustering [6] and ridge estimation [11].

11.2 Methodology

11.2.1 The Standard Approach

Let us first introduce the problem of nonparametric maximum likelihood estimation in the univariate case, proposed for the first time in [13]. Let $X_1, \ldots, X_n$ be i.i.d. observations with distribution function F and density f on a bounded domain $\Omega \subset \mathbb{R}$. The idea is to maximize a functional

$$L(f) - \lambda R(f) \tag{11.1}$$

where $L(f) = \sum_i \log f(x_i)$ is the log-likelihood, $R(f)$ is the roughness penalty, and the parameter $\lambda > 0$ controls the amount of smoothness. The penalty $R(f)$ is necessary to have a well defined likelihood, that would otherwise be unbounded because of the infinite class of functions we are considering. The idea is to reduce the space of possible solution in order to avoid trivial solutions such as the sum of delta functions at the observations. This can be achieved by penalizing too rough estimates. To measure the roughness or complexity of the estimate, in [13] the authors consider the functional $R(f) = ||(\sqrt{f})^{(1)}||_2^2$. Further developments of this model are presented in [23], where the authors consider a regularization functional of the form $R(f) = ||(\log f)^{(3)}||_2^2$.

11.2.2 Proposed Model and Estimation Procedure

We now consider the problem of estimating a density function f on a complex spatial domain. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d. observations drawn from a distribution F with density f on a bounded planar domain $\Omega \subset \mathbb{R}^2$. Likewise in in [23] we consider the logarithm tranform $g = \log(f)$, where g is a real function on Ω.

As discussed in the previous section, some type of regularization is necessary to avoid an unbounded likelihood. Here we consider a regularization functional of the form

$$R(g) = \int_\Omega (\Delta g)^2 \, dx \quad \text{where} \quad \Delta g = \frac{\partial^2 g}{\partial x_1^2} + \frac{\partial^2 g}{\partial x_2^2} \, .$$

where $\boldsymbol{x} = (x_1, x_2)$. The functional Δg is called Laplacian, and represents a measure of local curvature of g. It therefore controls the smoothness of the estimates while reducing the space of possible solutions. A key feature of the Laplacian is the invariance with respect to Euclidean transformations of the spatial coordinates. It therefore ensures that the concept of smoothness does not depend on the orientation of the coordinate system. Under appropriate boundary conditions, the density corresponding to the null family of the operator, i.e. when $\lambda \to +\infty$, is the uniform ditribution over Ω.

From a theoretical perspective, an analogous regularized nonparametric likelihood approach has been considered in the context of simple multidimensional domains in [14], using spline basis. The authors develop an elegant theoretical framework to study the asymptotic properties of such penalized density estimators. The generalization to multivariate domains with complex shapes is nonetheless not obvious. The main problems rely on the form of the regularizing functional and the discretization used, based on splines.

Here we propose a novel solution that exploit advanced numerical techniques, such as the finite element method. At first, we consider an appropriate discretization of the domain Ω. Since we are dealing with bounded domains, we can use constrained triangulations (see Figure 11.1). We then define a piecewise polynomial function over the discretized domain. In particular, let $\boldsymbol{\psi} := (\psi_1, \dots, \psi_K)^\top$ be the vector of linear finite element basis associated with the triangulation. Such basis are locally supported piecewise linear functions. We can define the discretized version of the function g as $\mathbf{g}^\top \boldsymbol{\psi}(x)$, where $\mathbf{g} \in \mathbb{R}^K$ is the vector of coefficients of the basis expansion. The penalization functional can be discretized by the quadratic form $\mathbf{g}^\top R_1 R_0^{-1} R_1 \mathbf{g}$, with

$$R_0 = \int_\Omega (\boldsymbol{\psi}\boldsymbol{\psi}^\top) \qquad \text{and} \qquad R_1 = \int_\Omega (\boldsymbol{\psi}_{\boldsymbol{x}_1}\boldsymbol{\psi}_{\boldsymbol{x}_1}^\top + \boldsymbol{\psi}_{\boldsymbol{x}_2}\boldsymbol{\psi}_{\boldsymbol{x}_2}^\top)\,,$$

where $\boldsymbol{\psi}_{\boldsymbol{x}_1} = (\partial\psi_1/\partial x_1, \dots, \partial\psi_k/\partial x_1)^\top$ and $\boldsymbol{\psi}_{\boldsymbol{x}_2} = (\partial\psi_1/\partial x_2, \dots, \partial\psi_k/\partial x_2)^\top$. For details on the derivation of the discretized regularization functional, see for instance [22].

11.3 Future Research

In this section we discuss some possible extensions of the proposed density estimation method. A first possibility is to consider higher dimensional and non-euclidean domain. For example, two-dimensional surfaces or complex three-dimensional bounded regions. Modern applications often require the analysis of data observed over these complex domains (see, e.g., [20]). In neuroscience researches for instance, brain studies are carried out on the cerebral cortex, a two dimensional curved domain with an highly convoluted nature [16, 8], or consider the brain as a whole, a three-dimensional domain with highly complex internal and external boundaries. In other fields, such as geoscience, data are often distributed over bounded non-planar domains. Flexible density estimation methods are therefore required to overcome the classical concept of Euclidean distance. In the case of Riemaniann manifolds, some proposals based on exponential maps are offered by [17, 4]. The finite element formulation in the proposed framework gives enough flexibility for possible extensions to curved two-dimensional domains and to complex three-dimensional domains. In particular, we can resort to surface finite elements, as in [16], and to volumetric finite elements.

Another possibility is to develop time-dependent density estimators. This type of modelization allows for the study of the evolution of underlying processes generating the data. The topic has drawn very little attention, especially in more than one dimension (see, e.g., [12] and references therein, for some first proposals in this regard). In the proposed approach, the generalization might consider two regularizations, one in time and one in space, or alternatively a unique regularization involving a time-dependent differential operator, in analogy to the spatio-temporal regression methods presented in [3] and [1].

Finally, a fascinating alternative is to tell the whole story from a bayesian perspective. The penalization has indeed the form of a Gaussian prior over a graph, the triangulation. This may lead to interesting considerations in terms of random processes, especially in the case of Poisson intensity estimation.

References

[1] Arnone, E., Azzimonti, L., Nobile, F., Sangalli, L.M.: Modeling spatially dependent functional data via regression with differential regularization. Journal of Multivariate Analysis **170**, 275–295 (2019)

[2] Azzimonti, L., Nobile, F., Sangalli, L.M., Secchi, P.: Mixed finite elements for spatial regression with PDE penalization. SIAM/ASA Journal on Uncertainty Quantification **2**(1), 305–335 (2014)

[3] Bernardi, M.S., Sangalli, L.M., Mazza, G., Ramsay, J.O.: A penalized regression model for spatial functional data with application to the analysis of the production of waste in Venice province. Stochastic environmental research and risk assessment **31**(1), 23–38 (2017)

[4] Berry, T., Sauer, T.: Density estimation on manifolds with boundary. Computational Statistics & Data Analysis **107**, 1–17 (2017)

[5] Cule, M., Samworth, R., Stewart, M.: Maximum likelihood estimation of a multi-dimensional log-concave density. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **72**(5), 545–607 (2010)

[6] Chacón, J.E.: A population background for nonparametric density-based clustering. Statistical Science **30**(4), 518–532 (2015)

[7] Chen, Y.C., Ho, S., Freeman, P.E., Genovese, C.R., Wasserman, L.: Cosmic web reconstruction through density ridges: method and algorithm. Monthly Notices of the Royal Astronomical Society **454**(1), 1140–1156 (2015)

[8] Chung, M.K., Hanson, J.L., Pollak, S.D.: Statistical analysis on brain surfaces. Handbook of Neuroimaging Data Analysis **233** (2016)

[9] Delicado, P., Giraldo, R., Comas, C., Mateu, J.: Statistics for spatial functional data: some recent contributions. Environmetrics: The official journal of the International Environmetrics Society **21**(3–4), 224–239 (2010)

[10] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., Picard, D.: Density estimation by wavelet thresholding. The Annals of Statistics, 508–539 (1996)

[11] Genovese, C.R., Perone-Pacifico, M., Verdinelli, I., Wasserman, L.: Nonparametric ridge estimation. The Annals of Statistics **42**(4), 1511–1545 (2014)

[12] Gervini, D.: Doubly stochastic models for replicated spatio-temporal point processes. arXiv preprint:1903.09253 (2019)

[13] Good, I.J., Gaskins, R.A.: Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. Journal of the American Statistical Association **75**(369), 42–56 (1980)

[14] Gu, C., Qiu, C.: Smoothing spline density estimation: Theory. The Annals of Statistics, 217–234 (1993)

[15] Leonard, T.: Density estimation, stochastic processes and prior information. Journal of the Royal Statistical Society: Series B (Methodological) **40**(2), 113–132 (1978)

[16] Lila, E., Aston, J.A., Sangalli, L.M.: Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging. The Annals of Applied Statistics **10**(4), 1854–1879 (2016)

[17] Kim, Y.T., Park, H.S.: Geometric structures arising from kernel density estimation on Riemannian manifolds. Journal of Multivariate Analysis **114**, 112–126 (2013)

[18] Menafoglio, A., Gaetani, G., Secchi, P.: Random domain decompositions for object-oriented Kriging over complex domains. Stochastic environmental research and risk assessment **32**(12), 3421–3437 (2018)

[19] Jorge, M., Romano, E.: Advances in spatial functional statistics. Stochastic environmental research and risk assessment (2016)

[20] Niu, M., Cheung, P., Lin, L., Dai, Z., Lawrence, N., Dunson, D.: Intrinsic Gaussian processes on complex constrained domains. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **81**(3), 603–627 (2019)

[21] Ramsay, T.: Spline smoothing over difficult regions. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64**(2), 307–319 (2002)

[22] Sangalli, L.M., Ramsay, J.O., Ramsay, T.O.: Spatial spline regression models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **75**(4), 681–703 (2013)

[23] Silverman, B.W.: On the estimation of a probability density function by the maximum penalized likelihood method. The Annals of Statistics 795–810 (1982)

[24] Wahba, G.: Spline models for observational data (Vol. 59). Siam (1990)

[25] Wand, M.P., Jones, M.C.: Kernel smoothing. Chapman and Hall/CRC (1994)

[26] Wood, S.N., Bravington, M.V., Hedley, S.L.: Soap film smoothing. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70**(5), 931–955 (2008)

Chapter 12
A Conformal Approach for Distribution-free Prediction of Functional Data

Matteo Fontana, Simone Vantini, Massimo Tavoni and Alexander Gammerman

Abstract Interval prediction has always been a complex problem to solve in the realm of Functional Data Analysis, and the solutions currently proposed to address this very important theoretical and applied issue are not satisfactory. In this contribution we propose a novel approach, based on a non-parametric forecasting approach coming from machine learning, called Conformal Prediction. In the scalar setting, the method is based on simple yet remarkable considerations about sample quantiles. After having stated in a formal way the issue of forecasting for functional data, we develop an algorithm that can be used to generate non-parametric prediction bands for a functional-on-scalar linear regression model. These forecasts are proven to be valid in a statistical sense (i.e., they guarantee a global coverage probability larger or equal to a given threshold) under a very minimal set of assumptions, and thus extremely useful in the statistical practice. The method is then tested on a real world application, namely ensemble emulations for climate economy models, very used in the climate change economics realm.

Matteo Fontana (✉)
MOX - Department of Mathematics, Politecnico di Milano, Italy,
e-mail: matteo.fontana@polimi.it

Simone Vantini
MOX - Department of Mathematics, Politecnico di Milano, Italy,
e-mail: simone.vantini@polimi.it

Massimo Tavoni
Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Italy
RFF-CMCC European Institute on Economics and the Environment (EIEE), Fondazione CMCC,
Lecce, Italy, e-mail: massimo.tavoni@polimi.it

Alexander Gammerman
Computer Learning Research Center - Department of Computer Science, Royal Holloway,
University of London, UK, e-mail: a.gammerman@rhul.ac.uk

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_12

12.1 Introduction

Functional Data Analysis (FDA) [16, 3] is a vibrant and dynamic field of statistics, that aims to develop theory and methods dealing with data points that come in the form of uni- or multivariate functions defined over a one or multi-dimensional domain. The typical mathematical setting of functional data consists in a random sample of independent real-valued univariate functions $Y_1(t), \ldots, Y_n(t)$ defined over the compact interval $T = [t_0, t_1] \subset \mathbb{R}$. The most common embedding of such functions is a Hilbert space such as L^2, but other embeddings are possible, such as Sobolev spaces ([18]) to deal with differential information or Bayes spaces for functional compositions [6].

Still open in FDA research is the the definition and computation of meaningful prediction bands. despite some attempts relying on a parametric setting have been performed [7, 2], it can be argued that the distributional assumptions at the core of these specific methods can't be tested in the statistical practice, which makes these forecasting methods not useful for many real world applications.

The main idea of the present work is to use a relatively new prediction framework, called Conformal Prediction (CP) [19], to develop a statistical method able to produce global prediction bands for functional data, while having guarantees on their statistical validity. Conformal Prediction was developed in the Machine Learning community as a method to define prediction intervals for Support Vector Machines [5]. It has been subsequently used by statisticians as a way to develop distribution-free prediction intervals for regression, both in the low [9] and high-dimensional [10] setting.

This contribution is structured as follows: in Section 12.2 we extend the CP framework to Functional Data Analysis, while in Section 12.3 we discuss the specific role of the Conformity Measure (CM) in predicting functional data in a conformal way, and how CMs define the shape of the prediction set calculated using CP. We then provide an application to a problem arising in climate change economics in Section 12.4.

12.2 Conformal Prediction for Functional Variables

Let us consider a regressive framework, in which we observe n data points

$$Z_1, \ldots, Z_n \sim P$$

where each Z_i, $i = 1, \ldots, n$ is IID. Each Z_i is a tuple $\{X_i, Y_i\}$ in $\mathbb{R}^p \times L^2$. Y_i is the response, a function embedded in $L^2[T]$ space, that is the space of L^2 functions defined over the interval $T = [t_0, t_1]$. X_i is instead a $p-$dimensional vector of covariates $X_i = \left[X_{i,1}, \ldots, X_{p,1}\right]$.

The objective is predicting with confidence a new (functional) value Y_{n+1} from a new vector of covariates X_{n+1} by using a regression operator of the form

$$\mu(x) = \mathbb{E}\left(Y | X = x\right), \, Y \in L^2[T], \, x \in \mathbb{R}^p. \tag{12.1}$$

In a more formal way, given a nominal type-1 error level $\alpha \in (0, 1)$, the idea is to identify sets C in $L^2[T]$ for which the following property holds:

$$\mathbb{P}\left(Y_{n+1} \in C(X_{n+1})\right) \geq 1 - \alpha \tag{12.2}$$

Please note that our objective is more ambitious than the one described in [8], where the authors aim to identify a prediction set not for Y_{n+1} itself, but for $\Pi\left(Y_{n+1}\right)$, its projection over a basis of $L^2[T]$.

The driving example in our case is Linear Function-on-Scalar regression , a specific case of a Functional Linear Model (FLM) [16] in which the covariates are scalars, the response is a function, and $\mu(x)$ is approximated using a t-dependent linear combination of the covariates x. Please note that in the specified setting we do not require μ to be the exact conditional mean: all the results presented still stand also in the case of poorly estimated or even mis-specified regression model.

A Conformal Prediction framework is quite apt to address the problem of identification of prediction sets as described by Equation 12.2. Moreover, the use of CP requires a very minimal set of assumptions (namely, the data tuples being IID).

The basic reasoning behind the developement of CP is due to an intuitive yet remarkable result about the probabilistic properties of sample quantiles. Let $U_1, \ldots, U_n$ be IID samples of a random scalar variable[1]. For a given level α and after having independentely sampled U_{n+1} from the same distribution, it can be noted that

$$\mathbb{P}\left(U_{n+1} \leq \hat{q}_{1-\alpha}\right) \geq 1 - \alpha \tag{12.3}$$

where $\hat{q}_{1-\alpha}$ is the sample quantile defined as

$$\hat{q}_{1-\alpha} = \begin{cases} U_{(\lceil (n+1)(1-\alpha) \rceil)}, & \text{if } \lceil (n+1)(1-\alpha) \rceil \leq n \\ +\infty & \text{otherwise.} \end{cases} \tag{12.4}$$

where $U_{(1)}, \ldots, U_{(n)}$ is a order statistic of $U_1, \ldots, U_n$. The previous result is valid not only in the univariate setting, where the order statistic is of straightforward definition, but also in the multivariate and functional case, after having defined a suitable notion of ordering. The statement in 12.3 is easy to verify: If $U_1, \ldots, U_n$ are exchangeable, the rank of U_{n+1} among the previous observations will be distributed as a discrete uniform over $\{1, \ldots, n, n+1\}$. As shown by [10], starting from this simple result, one can build with relative ease prediction sets for linear regression models, with the only assumption of the data being IID and that the regression operator being invariant to permutations of the training data.

We now provide an extension of the framework in [10] to the functional case. Let $y \in L^2$ be a new (trial) value. For each y we train the regression estimator $\hat{\mu}_y$ on the dataset $Z_1, \ldots, Z_n, (X_{n+1}, y)$. We then define the functional regression residuals

[1] Please note that the IID assumption is actually too strict, and the whole argument stands also under the weaker notion of exchangeability

$$\begin{aligned} E_{y,i} &= \left(Y_i - \hat{\mu}_y(X_i)\right),\ i = 1,\dots,n \\ E_{y,n+1} &= \left(y - \hat{\mu}_y(X_{n+1})\right). \end{aligned} \tag{12.5}$$

Please note that, being Y_i and y objects in L^2, also the residuals are actually residual functions, not scalars. For this reason, the ranking step proposed at this point in [10] is now non-trivial, and its choice will have important consequences on the interpretability of the forecast bands provided. Let $\mathcal{R}\left(\left\{E_{y,1},\dots,E_{y,n},E_{y,n+1}\right\}\right) : (L^2[T])^{n+1} \to \mathbb{R}^{n+1}$ be a functional able to provide a scoring criterion among $E_{y,1},\dots,E_{y,n},E_{y,n+1}$, and $\rho(E_{y,i})$ the value of this score for the function $E_{y,i}$. Let us assume that the criterion behaves like a depth measure, assigning high values to points that are deep in the point cloud, and low values to shallow points. We can then build a statistic $\pi(y)$ by computing

$$\pi(y) = \frac{1}{n+1}\sum_{i=1}^{n+1} 1\left\{\rho(E_{y,i}) \le \rho(E_{y,n+1})\right\} \tag{12.6}$$

where $1\{\cdot\}$ is the indicator function. Essentially, $\pi(y)$ represents the proportion of fitted residuals in the augmented sample that have a lower score than the last one $R_{y,n+1}$. By exchangeability of $Z_1,\dots,Z_{n+1}$ and the permutational invariance of $\hat{\mu}$ we can see that $\pi(Y_{n+1}$ is distributed as a uniform defined over the set $\left\{\frac{1}{n+1},\frac{2}{n+1},\dots,1\right\}$. This means that

$$\mathbb{P}\left((n+1)\pi(Y_{n+1}) \le \lceil(1-\alpha)(n+1)\rceil\right) \ge 1-\alpha \tag{12.7}$$

which can be alternatively interpreted by saying that $1-\pi(Y_{n+1})$ is a valid (conservative) $p-$value to test the null hypothesis $H_0 : Y_{n+1} = y$. By calculating such $p-$value over all possible values of $y \in L^2$ and then perform a thresholding to the desired level, we can define in a fairly straightforward way our conformal prediction interval for X_{n+1} in the following way.

$$C_{conformal}(X_{n+1}) = \left\{y \in L^2 : (n+1)\pi(Y_{n+1}) \le \lceil(1-\alpha)(n+1)\rceil\right\} \tag{12.8}$$

The whole procedure is summarised in Algorithm 1 of [4]. [22] prove that conformal prediction sets defined as in 12.8 have valid finite-sample coverage: in fact, Vovk et al. proofs are valid for very general covariates and responses (i.e. elements of measurable sets), much more general than the relatively well-behaved L^2 functions.

While being theoretically valid, this version of Conformal Prediction is very hard and computationally intensive to be implemented: As noted by [22], the "search" step has to be performed on a regular grid on the set in which the $Y_i, i = 1,\dots,n$ are embedded. Defining a regular grid in L^2 is a nontrivial task, and will reasonably yield a grid of points of such a size that will render any real-world implementation of the full version of CP non-feasible. For this reason, instead of the Full or Transductive Conformal Prediction framework, we will use the Split [8] or Inductive [15, 22] Conformal Prediction that is able to ease, by a smarter use of the training data, the computational burden associated to CP. The essential idea behind Split Conformal

Prediction is, as the name suggests, random splitting the training data in two samples, a proper training one and a calibration one. We present the Split Conformal prediction version of the method in Algorithm 2 of [4].

12.3 Functional Conformity Measures

To perform a Conformal Prediction task we need to choose a method to order residual curves. While for data lying in $\mathbb{R}$ this issue is straightforward, this is no longer the case even on the Euclidean plane $\mathbb{R}^2$. The most common idea to perform such ordering is the use of data depth, which allows also generalizations of univariate concepts such as medians and quantiles in the multivariate setting ([11] and references therein).

While many choices for a functional conformity or non-conformity measures can be performed, such as the use of small-ball probabilities ([3], specifically Chapter 4 and 13) used as a proxy for density in the L^2 case, we are focusing our attention to depth measures for functional data, and specifically the Band Depth ([12]).

As shown by Lopez-Pintado and Romo in [12] and by [20], the concept of depth in the functional setting can be effectively used to develop functional versions of univariate nonparametric techniques and visualization methods such as rank tests and boxplots. In fact, Band Depth is particularly apt to be used for identifying α-regions (in the sense of [11]) in the shape of bands. and thus very effective in reaching the goal of identifying prediction bands for functional data. Additional information about the choice of band depth can be found in [4]. The concept of band depth can be then effectively used as the ordering criterion in CP: we present the specific version of the algorithm in [4]

We have argued in favour of the validity of set predictions calculated through CP methods: it should also be noted that the bands identified using CP are valid also in a global sense, meaning that its coverage properties are valid for $\forall t \in T$. The proof, along some considerations about the pratical identification of the prediction set are presented in [4]

12.4 Application to IAM Ensemble Forecasting

As a test case for this newly developed method, we present an application to climate change economics, in which we create a statistically valid scenario and policy emulator via the use of functional on scalar linear models and the forecasting methods described in Section 12.3.

Climate Change is by far, the greatest policy challenge the humankind is facing: according to the last Intergovernmental Panel for Climate Change (IPCC) report [1] more decisive actions must be undertaken now, if we want to contain the average increase of the World surface temperature by 1.5°C, and avoid severe disruptions to the earth climate system. A fundamental tool to understand and explore the complex dynamics that regulates this phenomenon is the use of computer models, such as Integrated Assessment Models [14]. By integrating an economic and a climatic module, these models are able to simulate the profile of a variable of interest on a given timescale (usually CO_2 over the next century). Predicting a quantity for such a long time scale is a notoriously hard task, with a great degree of uncertainty involved. Many efforts have been undertaken to model and control this uncertainty, such as the development of standardized scenarios of future developement, called Shared Socioeconomic Pathways (SSPs) [21, 17] or the use of model ensembles to tackle the issue of model uncertainty. All the details about the application can be found in [4], while an example of prediction bands generated using the described version of CP for functional data can be seen in Figure 12.1

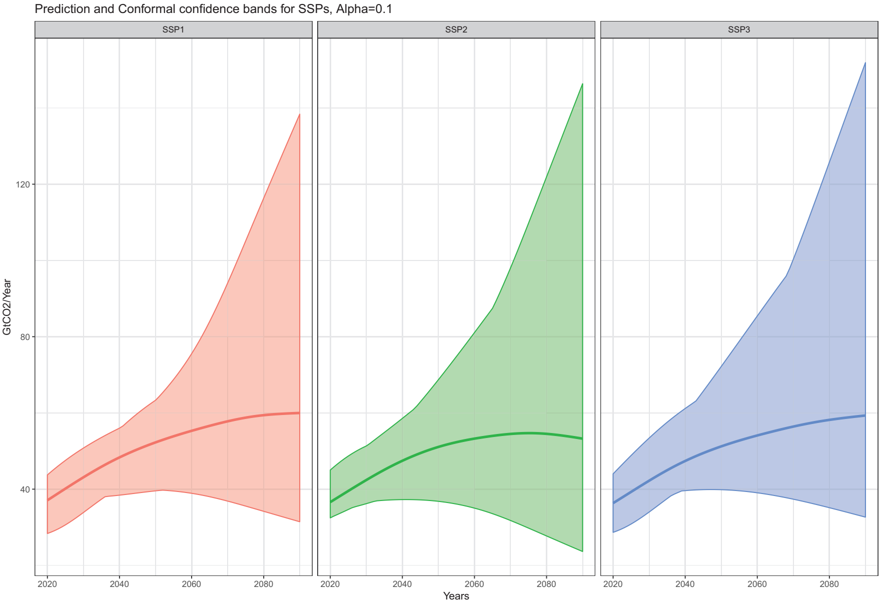


Fig. 12.1 Prediction Bands for SSP1, SSP2 and SSP3, $\alpha = 0.10$

References

[1] IPCC: Global warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty. In Press (2018)

[2] Antoniadis, A., Brossat. X., Cugliari, J., Poggi, J.M.: A prediction interval for a function-valued forecast model: Application to load forecasting. International Journal of Forecasting **32**(3), 939–947 (2016)

[3] Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis: Theory and Practice. New York, Springer (2007)

[4] Fontana, M., Vantini, S., Tavoni, M., Gammerman, A., Vovk, V.: Distribution-Free Conformal Prediction of Functional Data, with an application to climate economy model interval forecasting. MOX Report, Politecnico di Milano (2019)

[5] Gammerman, A., Vovk, V., Vapnik, V.: Learning by Transduction. In: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. UAI'98, pp. 148–155. San Francisco, Morgan Kaufmann Publishers Inc. (1998)

[6] Hron, K., Menafoglio, A., Templ, M., Hrůzová, K., Filzmoser, P.: Simplicial principal component analysis for density functions in Bayes spaces. Computational Statistics & Data Analysis **94**, 330–350 (2016)

[7] Hyndman, R.J., Shahid Ullah, M.: Robust forecasting of mortality and fertility rates: A functional data approach. Computational Statistics & Data Analysis **51**(10), 4942–4956 (2007)

[8] Lei, J., Rinaldo, A., Wasserman, L.: A conformal prediction approach to explore functional data. Annals of Mathematics and Artificial Intelligence **74**(1), 29–43. (2015)

[9] Lei, J., Robins, J., Wasserman, L.: Distribution-Free Prediction Sets. Journal of the American Statistical Association **108**(501), 278–287 (2013)

[10] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L.: Distribution-Free Predictive Inference for Regression. Journal of the American Statistical Association **113**(523), 1094–1111 (2018)

[11] Liu, R.Y., Parelius, J.M., Singh, K.: Multivariate analysis by data depth: descriptive statistics, graphics and inference. The Annals of Statistics **27**(3), 783–858 (1999)

[12] López-Pintado, S., Romo, J.: On the Concept of Depth for Functional Data. Journal of the American Statistical Association **104**(486), 718–734 (2009)

[13] Marangoni, G., Tavoni, M., Bosetti, V., Borgonovo, E., Capros, P., Fricko, O., et al.: Sensitivity of projected long-term CO2 emissions across the Shared Socioeconomic Pathways. Nature Climate Change **7**(2), 113–117 (2017)

[14] Nordhaus, W.D.: Rolling the 'DICE': an optimal transition path for controlling greenhouse gases, Resource and Energy Economics. **15**(1), 27–50 (1993)

[15] Papadopoulos, H., Proedrou, K., Vovk, V., Gammerman, A.: Inductive Confidence Machines for Regression. In: Machine Learning: ECML 2002. Lecture Notes in Computer Science, pp. 345–356. Springer, Berlin, Heidelberg (2002)

[16] Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. 2nd ed. Springer series in statistics. New York (2005)

[17] Riahi, K. et al.: The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview. Global Environmental Change **41**, 153–168 (2017)

[18] Sangalli, L.M., Secchi, P., Vantini, S., Veneziani, A.: A Case Study in Exploratory Functional Data Analysis: Geometrical Features of the Internal Carotid Artery. Journal of the American Statistical Association **104**(485), 37–48 (2009)

[19] Shafer, G., Vovk, V.: A Tutorial on Conformal Prediction. J Mach Learn Res **9**, 371–421 (2008)

[20] Sun, Y., Genton, M.G.: Functional Boxplots. Journal of Computational and Graphical Statistics **20**(2), 316–334 (2011)

[21] van Vuuren, D.P., Edmonds, J., Kainuma, M. et al.: The representative concentration pathways: an overview. Climatic Change **109**, 5 (2011)

[22] Vovk, V., Gammerman, A., Shafer, G.: Algorithmic learning in a random world. New York, Springer (2005)

Chapter 13
G-Lasso Network Analysis for Functional Data

Lara Fontanella, Sara Fontanella, Rosaria Ignaccolo, Luigi Ippoliti and Pasquale Valentini

Abstract Network analytical tools are becoming increasingly popular in analysing interdependent and interacting data entities. Statistical modelling of network data seeks to recover the underlying relational structure of the data capturing relevant characteristics and regularities in the pattern of interactions. This framework is widely adopted in multivariate data setting. However, in many applications, data are naturally regarded as random functions rather than multivariate vectors. In this work, we propose a simple approach to extend network analytical tools to the functional data setting. Specifically, we show that the graph representation of a set of functions can be retrieved through the precision matrix of a Gaussian Process, which encodes the conditional dependence structure among functional data. By using the standard graphical Lasso algorithm, preliminary results of the proposed methodology are shown for a benchmark dataset of daily average temperatures.

13.1 Introduction

Network science is a modern discipline that has gained popularity in many scientific fields, such as social sciences, medicine, physics and environmental science. Any set of data consisting of different entities with a relational structure can be described as

Lara Fontanella
University of Chieti-Pescara, Pescara, Italy, e-mail: lfontan@unich.it

Sara Fontanella (✉)
University of Torino, Torino, Italy and Imperial College London, UK,
e-mail: s.fontanella@imperial.ac.uk,

Rosaria Ignaccolo
University of Torino, Torino, Italy, e-mail: rosaria.ignaccolo@unito.it

Luigi Ippoliti
University of Chieti-Pescara, Pescara, Italy, e-mail: ippoliti@unich.it

Pasquale Valentini
University of Chieti-Pescara, Pescara, Italy, e-mail: pvalent@unich.it

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_13

a network. Statistical modelling of this data enables the relational structures between the entities to be evaluated, as well as enabling the identification both of the most influential objects and of those groups of nodes, known as communities, that are internally well connected. This analytical approach is popular because of its flexibility and because it enables network connectivity structures and measures to be conceptualized. Graph theory, in this context, can transform these complex systems into useful mathematical representations. Probabilistic graphical models, in which it is assumed that the inferred graph structure encodes the conditional dependence relationship among random variables represented by the data entities, is one useful approach for graph-learning. Although this framework is well-established for vector-valued data, little attention has been given to functional data, and effective statistical modelling of network functional data is still under development and requires innovative theories. The literature on functional graphical models is still sparse with only a few recent papers available [24, 17, 18].

With the advance of modern technology, huge amounts of data are being recorded continuously over some intervals or intermittently at several discrete points along a domain that can be time, but also depth or altitude for example. These are examples of functional data, which are intrinsically infinite-dimensional. Instead of considering their discrete representation, in Functional Data Analysis (FDA) curves are treated as single entities (for an overview see [19, 8]).

In this work, we evaluate the usefulness of network analytical models applied to the functional data domain. In particular, we consider network functional data to be modelled through undirected graphs, where the vertices represent random functions. The graph is learned through probabilistic graphical models and the inferred network encodes the conditional dependence structure among the functions. In this framework, we show how complex connections among functional data can be structured and made interpretable. Moreover, the importance of each function within the network can be studied through centrality analysis [10, 11, 4], which quantifies the ability of a function to influence other functions using its connection topology. The topological structure can also facilitate the identification of communities. Community detection [14], or network clustering, is a crucial step in the investigation of complex network structures. In this context, it involves the identification of groups of highly connected functions. This analysis may offer insights on how the network is organised and can facilitate classification of the functions, based on their role with respect to the communities to which they belong.

Preliminary results from network analysis of functional data are shown for the benchmark dataset of Canadian weather-stations [19].

13.2 Graphical Models for Functional Data

Let $\mathbf{Y} = \{Y_j\}_{j=1}^p$ be a collection of p square integrable random functions, where each component Y_j is defined in $L_2(\mathcal{T})$, and $\mathcal{T}$ is a compact subset of $\mathbb{R}$ [8]. Furthermore, suppose that Y_j is a zero mean Gaussian Process with covariance function $E\left[Y_j(t)Y_j(t')\right] = K_j(t,t')$, with $t, t' \in \mathcal{T}$. Then, it is well known [8] that there exist constants, $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$, together with continuous orthonormal basis functions, $\psi_1(\theta), \psi_2(\theta), \ldots$, for which Mercer's Theorem [21] provides the following convenient spectral decomposition

$$K_j(t,t') = \sum_{v=1}^{\infty} \lambda_v \psi_v(t) \psi_v(t'), \quad j = 1, \ldots, p$$

Then, using the Karhunen-Loéve expansion (KLE) we may represent $Y_j(t)$ as

$$Y_j(t) = \sum_{v=1}^{\infty} \alpha_v^{(j)} \ \psi_v(t) \tag{13.1}$$

with

$$\alpha_v^{(j)} = \int_{\mathcal{T}} Y_j(t) \psi_v(t) dt, \tag{13.2}$$

being zero mean uncorrelated random variables with variance λ_v.

13.2.1 Functional Gaussian Graphical Models

To evaluate the relationships among the functional data, the proposed model exploits the theory of graphical models to infer an undirected graph whose vertices are defined by the p random functions.

In particular, we assume an undirected graph $G = (V, E)$ to describe the conditional independence relationships of components in $\mathbf{Y}$, whereby G can be inferred through the graphical Lasso (G-Lasso) technique [23, 12]. An undirected graph $G = (V, E)$ is defined by a finite set of vertices $V = \{1, \ldots, p\}$, and a set of undirected edges $E \subseteq \{(i, j) : i, j \in V \times V, \ i \neq j\}$ that are encoded as pairs of vertices of G. Here, we assume that the functional variables $\{Y_1, Y_2, \ldots, Y_p\}$ belong to the undirected graph G and the absence of an edge between i and j, with $i, j = 1, \ldots, p$ and $i \neq j$, corresponds to the conditional independence of the two random functions given the remaining ones, i.e. $Y_i \perp Y_j | Y_{V \setminus i,j}$. This is known as the pairwise Markov property relative to G, which implies both the local and the global Markov properties relative to G [16]. Then, if the functional variables $\{Y_1, Y_2, \ldots, Y_p\}$ are from a p-dimensional Gaussian Process with covariance matrix $\mathbf{\Sigma}$, it turns out that every pair of functions not contained in the edge set is conditionally independent, given all remaining functions, and corresponds to a zero entry in the precision matrix $\mathbf{\Phi} = \mathbf{\Sigma}^{-1}$ [16].

13.2.2 Estimation

The pattern of zero entries in the elements ϕ_{ij} of the precision matrix $\mathbf{\Phi}$, which corresponds to conditional independence restrictions between functions, can be learned by using the G-Lasso algorithm. In particular, to achieve sparsity in the precision matrix, the algorithm imposes a L_1 penalty for the estimation of $\mathbf{\Phi}$ and maximises the log-likelihood

$$\log \det \, \mathbf{\Phi} - \mathrm{Trace}\Big(\mathbf{S\Phi}\Big) - \gamma \; ||\mathbf{\Phi}||_1$$

where $||\mathbf{\Phi}||_1$ is the L_1 norm (i.e., the sum of the absolute values of the elements of $\mathbf{\Phi}$) and

$$\mathbf{S} = \frac{1}{Nr} \sum_{l=1}^{N} \mathbf{A}_l' \mathbf{A}_l, \quad l = 1, \ldots, N$$

is the sample covariance matrix of expansion coefficients $\mathbf{A}_l = \big(\hat{\alpha}_l^{(1)} \; \hat{\alpha}_l^{(2)} \ldots \hat{\alpha}_l^{(p)}\big) \in \mathbb{R}^{r \times p}$ with $\hat{\alpha}_l^{(j)} = \big(\hat{\alpha}_{l1}^{(j)} \; \hat{\alpha}_{l2}^{(j)} \ldots \hat{\alpha}_{lr}^{(j)}\big)'$ being the vector of the first r estimated expansion coefficients of equation (13.2) associated with the l-th observed replication, $l = 1, \ldots, N$, of collection $\mathbf{Y} = \big\{Y_j\big\}_{j=1}^{p}$.

The L_1 penalized maximum likelihood estimator requires the selection of the regularization parameter γ, which directly controls the sparsity level of the graph G. When $\gamma = 0$ no penalty is imposed. Increasing γ results in an increasing number of zero entries in the precision matrix. The choice of the parameter γ is then crucial for retrieving the underlying network structure. The optimal computational properties of the G-Lasso algorithm facilitate the use of data-driven selection approaches. A common practice is that of estimating several graph structures, $G(\gamma)$, over a grid of values for γ, ranging from small to large. Selection criteria, such as Akaike Information Criterion (AIC) [1], Bayesian Information Criterion (BIC) [22], Extended Bayesian Information Criterion (EBIC) [9] can be then adopted to select the best model.

13.3 Network Analysis on the Canadian Weather

In this application, the Canadian temperature data are used as test set to evaluate network modelling approaches to functional data. The dataset consists of daily measured temperatures at 35 different locations in Canada over 365 days (see, e.g., [19], for more details)[1].

Since for each station only one functional observation is available (i.e. $N = 1$), each temperature profile was represented by the first 200 expansion coefficients by means of principal cubic splines [15].

As described in Section 13.2.2, to evaluate the relationships among the 35 weather stations, an undirected graph was then obtained from the precision matrix estimated through the G-Lasso algorithm[2]. The G-Lasso algorithm was run for 100 different values of the regularization parameter γ and the best model was selected using the EBIC index [6, 9].

To characterise the network, the centrality indices were also computed [14]. These measures provide valuable information on the importance and position of each function within the network, namely *strength*, *closeness*, and *betweenness*. The strength is a measure of direct connections of a function to the others and it is

[1] All statistical analyses were run in the programming language R. The data are available in the *fda* package [20].

[2] We used *qgraph* [7], *bootnet* [6] and *igraph* [5] packages for network estimation and visualization.

computed by summing the absolute value of all edge weights for a function with other functions in the network. Closeness is based on the length of the average shortest path between a function and all functions in the network and expresses how well the function is indirectly connected to the others. The betweenness measures how frequently the function lies on the shortest path between two other functions and identifies the nodes that act as bridges between the other functions in a network [7].

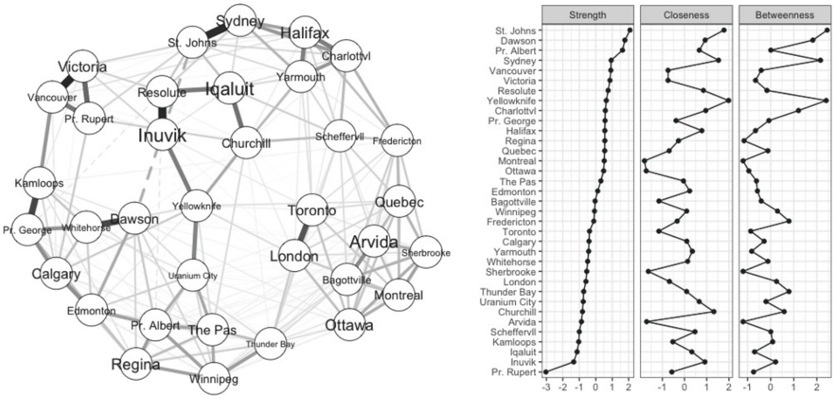


Fig. 13.1 Network of the Canadian weather dataset. Left panel: connectivity structure of the 35 Canadian weather stations. Right panel: centrality measures for each station.

Figure 13.1 shows the retrieved network, which is sparse due to the LASSO estimation: the network has only 217 non-zero edges out of 595 possible edges, with *density*, given by the number of existing relationships relative to the possible number of ties, equal to 0.364. The network structure seems to capture the geographical neighbouring relationship among the weather stations as strong connections emerge among the stations of Sidney and St. Johns, London and Toronto, Resolute and Inuvik, Dawson and Whitehorse. In general, close stations are highly connected, while station far apart do not show connections, meaning that their temperature profiles differ significantly.

The centrality indices show that functions differ quite substantially in their centrality estimates. In the network, St. Johns, Dawson and Pr. Albert have the highest strength, while Yellowknife has the highest betweenness and closeness. This indicates that while the formers are highly connected locally to their neighbours, sharing similar temperature profiles, Yellowknife is highly connected globally.

To further investigate the relationships among the weather stations, we evaluated the community structure of the networks derived from the empirical data, in order to identify groups of locations likely to be highly connected. We performed network clustering using the fast greedy modularity optimization algorithm for finding communities [3].

The community detection algorithm identified five clusters, which seem highly coherent with the actual geographical position of the stations (see Figure 13.2A). The red cluster consists of the temperature curves registered in the North Canada stations. The continental stations are grouped in the green cluster, while the orange group consists of the Pacific stations. Finally, the stations of the Atlantic coast are gathered in the purple and blue clusters. The latter is made up of the stations located in Quebec and Ontario, while the former primarily captures station situated in Nova Scotia. The obtained results are consistent with previous studies [2, 13].

The average temperature curves (Fig. 13.2C) clearly show that meteorological conditions are harsher in the northern part of Canada. Temperatures in the Pacific coast stations (orange cluster) are higher throughout the year and show low variability, while the continental cities have lower temperatures compare to the coast cities, especially during winter.

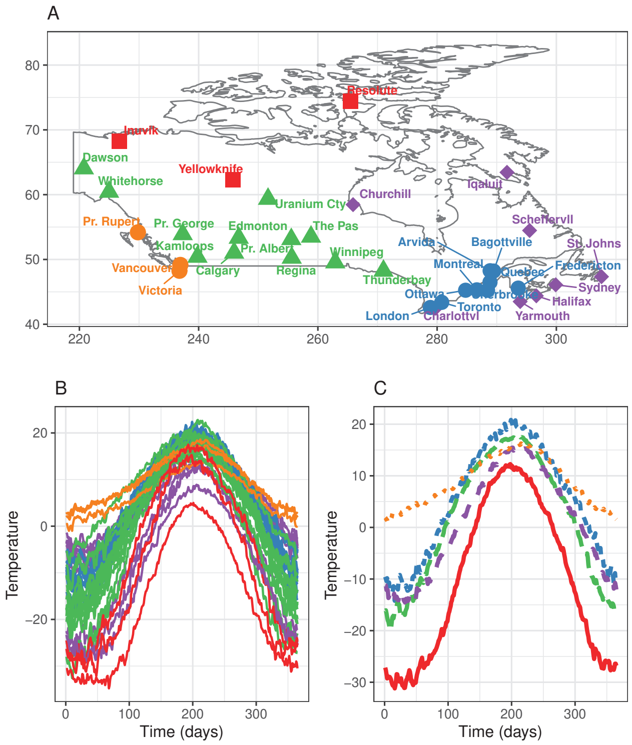


Fig. 13.2 A) Geographical position of the Canadian weather stations according to their cluster membership, B) Weather station temperature profiles stratified by clusters and C) Estimated mean temperature profile for each cluster.

13.4 Discussion

In this paper, we have considered the use of Gaussian graphical models for learning structures and graphs from functional data. For practical implementation, we have suggested the use of the simple G-Lasso algorithm. However, how to estimate the pattern of interactions among the functions is still an open problem of both theoretical and practical significance. Extensions of the present proposal will be an issue that we leave for future works.

References

[1] Akaike, H.: Information theory and an extension of the maximum likelihood principle, in B. N. Petrov & B. F. Csaki (Eds.), Second International Symposium on Information Theory, 267–281. Academiai Kiado, Budapest (1973)

[2] Bouveyron, C., Jacques, J.: Model-based clustering of time series in group-specific functional subspaces. Advances in Data Analysis and Classification **5**(4), 281–300 (2011)

[3] Clauset, A., Newman, M., Moore, C.: Finding community structure in very large networks. Phys. Rev. E **70**, 066111 (2004)

[4] Costa, L. d. F., Rodrigues, F. A., Travieso, G., Villas Boas, P. R.: Characterization of complex networks: A survey of measurements. Advances in Physics **56**(1), 167–242 (2007)

[5] Csardi, G., Nepusz, T.: The igraph software package for complex network research. InterJournal, Complex Systems, 1695 (2006)

[6] Epskamp, S., Borsboom, D., Fried, E.: Estimating psychological networks and their accuracy: A tutorial paper. Behavior Research Methods, **50**, 195–212 (2017)

[7] Epskamp, S., Cramer, A., Waldorp, L., Schmittmann, V., Borsboom, D.: qgraph: Network visualizations of relationships in psychometric data. Journal of Statistical Software **48**(4), 1–18 (2012)

[8] Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis: Theory and Practice (Springer Series in Statistics). Springer-Verlag, Berlin, Heidelberg (2006)

[9] Foygel, R., Drton, M.: Extended bayesian information criteria for gaussian graphical models. In: Lafferty, J.D., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A. (eds), Advances in Neural Information Processing Systems 23, 604–612. Curran Associates, Inc. (2010)

[10] Freeman, L.: A set of measures of centrality based on betweenness. Sociometry **40**(1), 35–41 (1977)

[11] Freeman, L.: Centrality in social networks conceptual clarification. Social Networks **1**(3), 215 – 239 (1978)

[12] Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical Lasso. Biostatistics **9**(3), 432–441 (2008)

[13] Jacques, J. Preda, C.: Model-based clustering for multivariate functional data. Computational Statistics & Data Analysis **71**, 92–106 (2014)

[14] Javed, M., Younis, M., Latif, S., Qadir, J., Baig, A.: Community detection in networks: A multidisciplinary review. Journal of Network and Computer Applications **108**, 87–111 (2018)

[15] Kent, J., Mardia, K.: Modelling strategies for spatial-temporal data. In Lawson, A. and Denison, D., editors, Spatial Cluster Modelling, 213–226. London: Chapman and Hall (2002)

[16] Lauritzen, S.: Graphical Models. Oxford Science Publications. Clarendon Press (1996)

[17] Li, B., Solea, E.: A nonparametric graphical model for functional data with application to brain networks based on fmri. Journal of the American Statistical Association **113**(524), 1637–1655 (2018)

[18] Qiao, X., Guo, S., James, G.M.: Functional graphical models. Journal of the American Statistical Association **114**(525), 211–222 (2019)

[19] Ramsay, J., Silverman, B.: Functional Data Analysis. Springer Series in Statistics. Springer (2005)

[20] Ramsay, J., Wickham, H., Graves, S., Hooker, G.: fda: Functional Data Analysis. R package version 2.4.8 (2018)

[21] Riesz, F., Sz-Nagy, B.: Functional analysis. Ungar, New York (1955)

[22] Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics **6**(2), 461–464 (1978)

[23] Yuan, M., Lin, Y.: Model selection and estimation in the Gaussian graphical model. Biometrika **94**(1), 19–35 (2007)

[24] Zhu, H., Strawn, N., Dunson, D.: Bayesian graphical models for multivariate functional data. Journal of Machine Learning Research **17**(204), 1–27 (2016)

Chapter 14
Modelling Functional Data with High-dimensional Error Structure

Yuan Gao, Han Lin Shang and Yanrong Yang

Abstract We propose to model raw functional data as a mixture of functions and high-dimensional error. The conventional approach to retrieve the functional component from raw data is through varied smoothing techniques. Nevertheless, smoothing itself may not be adequate when measurement error exists. We propose to use factor model to reduce the dimension of the high-dimensional measurement error, while smoothing the functional component. Our model also provides as an alternative for modelling functional data with step jump. Regularized least squares method is used to find the model estimates. We look at the asymptotic behaviour of the estimator when both the sample size and the number of points per curve go to infinity and the limiting distribution is derived.

14.1 Introduction

With the increasing capability of data storing, functional data analysis (FDA) has received growing attention in the last twenty years. Functional data are considered as realizations of smooth random objects, in graphical representations of curves, images, and shapes. The monographs of [12, 13] and [11] provide a comprehensive account of the methodology and applications of FDA. More recent advances in this field can be found in survey papers [3, 7, 8, 14, 17]. We denote a random sample of n functional data as $\mathcal{X}_i(u), i = 1, \dots, n$, and $u \in \mathcal{I}$, where $\mathcal{I}$ is a compact interval on the real line.

Yuan Gao (✉)
The The Australian National University, ACT 2601, Australia, e-mail: yuan.gao@anu.edu.au

Han Lin Shang
Macquarie University, Sydney, NSW 2109, Australia, e-mail: hanlin.shang@mq.edu.au

Yanrong Yang
the Australian National University, 26C Kingsley St. ACT, Australia,
e-mail: yanrong.yang@anu.edu.au

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_14

In practice, the observed data are discrete points and are often contaminated with noise or measurement error. If we use Y_{ij} to represent the jth observation on the ith subject, then the observed data can be expressed with the model

$$Y_{ij} = \mathcal{X}_i(u_j) + \eta_{ij}, \quad j = 1, \dots p.$$

We use $\mathcal{X}_i(u_j)$ to denote the realization of the jth discrete point on the curve $\mathcal{X}_i$, and η_{ij} is the measurement error. In this paper, we assume that measurement error only takes place where the measurements are taken. Thus the error η_{ij} is a multivariate term of dimension p. Even though in real data, the functional component $X_i(u_j)$ is of the same p dimension, it is different from η_{ij} in nature. We may impose smoothing assumptions on the functions, which usually means the function possesses one or more derivatives. This smoothness feature is used to separate the functions from the measurement errors. This is called functional smoothing.

Classic smoothing tools apply to functional data, including kernel methods [16]; local polynomial smoothing [6] and spline smoothing [15, 4, 9]. With pre-smoothed functions, estimates such as mean and covariance functions can be further obtained. Recent studies on functional smoothing approaches include [2, 18, 19].

However, smoothing tools alone may not be adequate in removing the error and may cause unstable estimation of the function in cases where systematic measurement error exist. In this context, the measurement error is more than pure white noise. The problem of measurement error arises in different fields such as survey data, nutrition data and environment studies. In order to model the measurement error term, we take a further look at η_{ij}. In FDA, it is often the case that the number of discrete points p on each subject is large compared to the sample size n. Hence the term η_{ij} is a high-dimensional component. This raises the problem of the curse of dimensionality, which naturally calls for dimension reduction models. Abundance studies have been conducted on various dimension reduction techniques on large data. Among them, factor models are widely used [1, 5, 10].

In this article, we propose to use a factor model on the measurement error term. The high-dimensional measurement error is assumed to be driven by a small number of latent factors.

$$\eta_{ij} = \boldsymbol{\lambda}_i^\top \boldsymbol{F}_j + \epsilon_{ij}, \qquad i = 1, \dots, n, j = 1, \dots, p,$$

where $F_j \in \mathbb{R}^r$ are the unobserved factors; $\lambda_i \in \mathbb{R}^r$ are the factor loadings and ϵ_{ij} are idiosyncratic errors with mean zero. Thus the observed data can be written as the sum of two components.

$$Y_{ij} = \mathcal{X}_i(u_j) + \boldsymbol{\lambda}_i^\top \boldsymbol{F}_j + \epsilon_{ij}, \qquad i = 1, \dots, n, j = 1, \dots, p.$$

Since we observe the data as a mixture, there is an identification problem between the two parts. It is required that the high-dimensional term η_{ij} is independent of the functional term $\mathcal{X}(u)$. Even though compared with simple smoothing model, our mixed model introduces more parameters due to the factor model assumed on the measurement error, it turns out that by simultaneously estimating the two parts, the total number of parameters is often equivalent or even smaller than the usual smoothing model itself.

14.2 Motivation

14.2.1 Functional Data with Measurement Error

Figure 14.1 shows the rainbow plots of the average daily temperature and log precipitation at 35 locations in Canada. Due to the nature of the two kinds of data, it is reasonable to assume that temperature and precipitation are functions over time. The two graphs, however, display distinct features. In the temperature plot, it is relatively easy to discern the shape of each curve, while in the precipitation plot, there is a great amount of variability in the raw data such that it is almost impossible to observe the underlying shape of the curves.

Smooth temperature data can be retrieved without much difficulty using basic smoothing techniques. The residuals are small with constant variation. On the other hand, for the precipitation data, the residuals after smoothing are of high variation and even contain some extreme values. It is reasonable to suspect the existence of measurement error. Our model endeavours to further explain the large residuals in similar cases as the precipitation data.

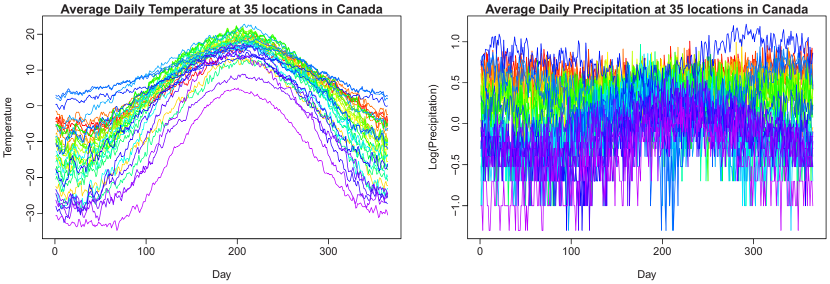


Fig. 14.1 Average daily temperature and log precipitation in 35 Canadian weather stations averaged over 1960 to 1994

14.2.2 Functional Data with Step Jump

We provide another example on functional data with step jump to motivate the proposed model. Suppose we observe a sample of raw functional data with step jump in the mean level as shown in Figure 14.2. The first graph shows the raw data, where a jump could be seen in the middle. The middle graph is smoothed functions generated by using B-spline smoothing with penalty. The residuals after smoothing is presented in the bottom graph. The large residuals around the jump make it clear that, the residuals still contain structures and without measures to deal with the step jump, smoothing itself is not enough to model this type of data. The proposed model applying to the same data generates smaller residuals and the model can be shown to be of less flexibility. By less flexibility, we mean that the model has fewer parameters or degrees of freedom. This is indeed one of the main goals in view of model selection methods.

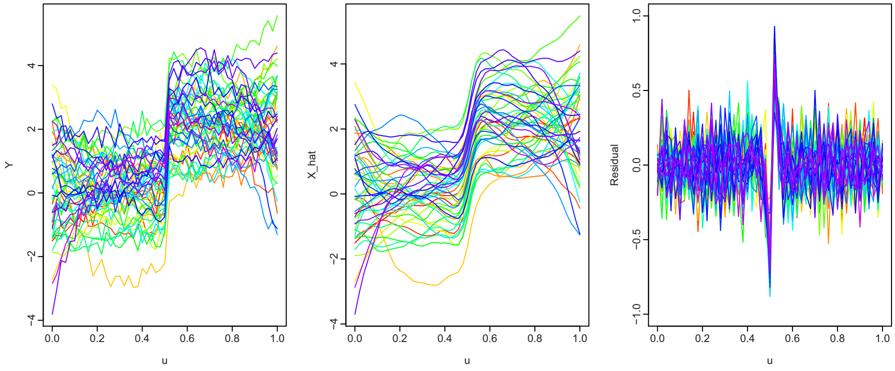


Fig. 14.2 Simulated sample of functional data with step jump

14.3 Model Specification and Estimation

14.3.1 Model with Basis Expansion

We consider a sample of functional data $\mathcal{X}_i(u), i = 1, \dots, n$, which takes values in the space $H := L^2(\mathcal{I})$ of real-valued square integrable functions on $\mathcal{I}$. The space H is a Hilbert space, equipped with the inner product $\langle x, y \rangle := \int x(u)y(u)du$. The function norm is defined as $\|x\| := \langle x, x \rangle^{1/2}$. The functional nature of $\mathcal{X}_i(u)$ allows us to represent it as a linear expansion of a set of smooth basis functions.

$$\mathcal{X}_i(u) = \sum_{k=1}^{K} c_{ik}\phi_k(u), \quad u \in \mathcal{I}$$

where $\phi_k(u)$ are a set of basis functions, and c_{ik} are the coefficient for the ith curve. Although functions are of infinite dimensionality, we regard $\mathcal{X}_i(u)$ as the target function that possesses the smoothing feature. This imposed smoothness condition implies we could write the function as a sum of finite basis functions. It does not mean that FDA simply reduces to multivariate data analysis and the number of K also depends on how the basis system is chosen. Therefore, we can write the full model as

$$Y_{ij} = \sum_{k=1}^{K} c_{ik}\phi_k(u_j) + \eta_{ij},$$
$$\eta_{ij} = \boldsymbol{\lambda}_i^\top \boldsymbol{F}_j + \epsilon_{ij}, \qquad i = 1, \dots, n, j = 1, \dots, p.$$

In this article, we treat the basis functions $\phi_k(u)$ as known. This is of course a simplification to accommodate for the theoretical proofs. In real data analysis, there exist a variety of choices for the basis functions and the decision can be quite subjective.

14.3.2 Estimation

We can write the model for the ith object as

$$\boldsymbol{Y}_i = \boldsymbol{\Phi c}_i + \boldsymbol{F\lambda}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \dots, n$$

Combining all the objects, we can write in matrix form

$$\boldsymbol{Y} = \boldsymbol{\Phi C} + \boldsymbol{F\Lambda}^\top + \boldsymbol{E},$$

where $\boldsymbol{Y}$ is $p \times n$ and $\boldsymbol{C}$ is a $K \times n$ matrix containing all coefficients. The matrix $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_n)^\top$ $n \times r$ and $\boldsymbol{E} = (\boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_n)$ is $n \times p$. Since $\boldsymbol{\Phi}$ is assumed to be known, we will illustrate how the parameters $\boldsymbol{C}, \boldsymbol{F}$ and $\boldsymbol{\Lambda}$ are estimated in the following.

First for the latent factor estimation, there is an identification problem such that $\boldsymbol{F\Lambda}^\top = \boldsymbol{FAA}^{-1}\boldsymbol{\Lambda}^\top$ for any $r \times r$ invertible matrix $\boldsymbol{A}$. Thus we impose the normalization restriction on the factor matrix $\boldsymbol{F}$ $\boldsymbol{F}^\top\boldsymbol{F}/p = \boldsymbol{I}_r$. It is also required for the factor loading matrix that $\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}$ is a diagonal matrix.

We propose to use a penalized least squares approach, where the objective function is defined as

$$SSR(\boldsymbol{c}_i, \boldsymbol{F}, \boldsymbol{\Lambda}) = \sum_{i=1}^{n} \left[(\boldsymbol{Y}_i - \boldsymbol{\Phi c}_i - \boldsymbol{F\lambda}_i)^\top (\boldsymbol{Y}_i - \boldsymbol{\Phi c}_i - \boldsymbol{F\lambda}_i) + \alpha PEN(\mathcal{X}_i) \right],$$

where $PEN(\mathcal{X}_i)$ is a penalty term used for regularization, and α is the tuning parameter controlling the degree of smoothness.To quantify the notion of "roughness" in a function, we use the square of the second derivative. Define the measure of roughness as $PEN_2(\mathcal{X}_i) = \int [D^2\mathcal{X}_i(s)]^2 ds$, where $D^2\mathcal{X}_i$ denotes taking the second derivative of the function $\mathcal{X}_i$. To simplify the penalty term denote

$$\boldsymbol{\Phi}(u) = (\phi_1(u), \dots, \phi_K(u))^\top. \tag{14.1}$$

Then $\mathcal{X}_i(u) = \boldsymbol{c}_i^\top \Phi(u)$. We can re-express the roughness penalty $PEN_2(\mathcal{X}_i)$ in a matrix form as the following:

$$
\begin{aligned}
PEN_2(\mathcal{X}_i) &= \int [D^2\mathcal{X}_i(s)]^2 ds = \int [D^2\boldsymbol{c}_i^\top \boldsymbol{\Phi}(s)]^2 ds \\
&= \int \boldsymbol{c}_i^\top D^2\boldsymbol{\Phi}(s) D^2\boldsymbol{\Phi}^\top(s)\boldsymbol{c}_i ds = \boldsymbol{c}_i^\top \left[\int D^2\boldsymbol{\Phi}(s) D^2\boldsymbol{\Phi}'(s) ds\right] \boldsymbol{c}_i \\
&= \boldsymbol{c}_i^\top \boldsymbol{R}\boldsymbol{c}_i
\end{aligned}
$$

where $\boldsymbol{R} = \int D^2\boldsymbol{\Phi}(s)D^2\boldsymbol{\Phi}'(s)ds$.

The penalty term is different for each subject only by the coefficient $\boldsymbol{c}_i$. Thus the objective function can be written as

$$
SSR(\boldsymbol{c}_i, \boldsymbol{F}, \boldsymbol{\Lambda}) = \sum_{i=1}^{n} \left[(\boldsymbol{Y}_i - \boldsymbol{\Phi}\boldsymbol{c}_i - \boldsymbol{F}\boldsymbol{\lambda}_i)^\top (\boldsymbol{Y}_i - \boldsymbol{\Phi}\boldsymbol{c}_i - \boldsymbol{F}\boldsymbol{\lambda}_i) + \alpha \boldsymbol{c}_i^\top \boldsymbol{R}\boldsymbol{c}_i \right],
$$

subject to the constraint $\boldsymbol{F}^\top \boldsymbol{F}/p = \boldsymbol{I}_r$.

To estimate the coefficient $\boldsymbol{c}_i$, define the projection matrix $\boldsymbol{M}_F = \boldsymbol{I}_p - \boldsymbol{F}(\boldsymbol{F}^\top\boldsymbol{F})^{-1}\boldsymbol{F}^\top = \boldsymbol{I}_p - \boldsymbol{F}\boldsymbol{F}^\top$. It is easy to obtain the standard least squares solution for the coefficient $\hat{\boldsymbol{c}}_i$, see the first equation in (14.2).

Next to estimate $\boldsymbol{F}$ and $\boldsymbol{\Lambda}$, we focus on the factor model term$\boldsymbol{\eta}_i = \boldsymbol{F}\boldsymbol{\lambda}_i + \boldsymbol{\epsilon}_i$, and in matrix form $\boldsymbol{Z} = \boldsymbol{F}\boldsymbol{\Lambda}^\top + \boldsymbol{E}$, where $\boldsymbol{Z} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n)$. The factor model is estimated using principal component. We perform eigen-decomposition on the covariance structure of the matrix $\boldsymbol{Z}$ to obtain the factor and the factor loadings. The covariance of the matrix $\boldsymbol{Z}$ can be calculated as $\boldsymbol{Z}\boldsymbol{Z}^\top = \sum_{i=1}^n \boldsymbol{\eta}_i\boldsymbol{\eta}_i^\top = \sum_{i=1}^n (\boldsymbol{Y}_i - \boldsymbol{\Phi}\boldsymbol{c}_i)(\boldsymbol{Y}_i - \boldsymbol{\Phi}\boldsymbol{c}_i)^\top$. It could be seen that $\boldsymbol{F}$ is needed to estimate $\boldsymbol{c}_i$, and in turn $\boldsymbol{c}_i$ is needed to estimate $\boldsymbol{F}$. The final estimator $(\widehat{\boldsymbol{c}}_i, \widehat{\boldsymbol{F}}, \widehat{\boldsymbol{\Lambda}})$ is the solution of the set of equations

$$
\begin{cases}
\widehat{\boldsymbol{c}_i} = \left(\boldsymbol{\Phi}^\top \boldsymbol{M}_{\hat{F}}\boldsymbol{\Phi} + \alpha\boldsymbol{R}^\top\right)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{M}_{\hat{F}}\boldsymbol{Y}_i, \qquad i = 1, \ldots, n \\
\left[\frac{1}{np}\sum_{i=1}^n (\boldsymbol{Y}_i - \boldsymbol{\Phi}\widehat{\boldsymbol{c}_i})(\boldsymbol{Y}_i - \boldsymbol{\Phi}\widehat{\boldsymbol{c}_i})^\top\right] \widehat{\boldsymbol{F}} = \widehat{\boldsymbol{F}}\boldsymbol{V}_{np},
\end{cases}
. \tag{14.2}
$$

We can estimate $\widehat{\boldsymbol{F}}$ and $\widehat{\boldsymbol{c}_i}$ using numerical iterations. The estimated factor loadings can be obtained by $\widehat{\boldsymbol{\Lambda}}^\top = \widehat{\boldsymbol{F}}^\top(\boldsymbol{Y} - \Phi\widehat{\boldsymbol{C}})$. Finally, the functional component can be estimated by $\widehat{\mathcal{X}_i}(u) = \widehat{\boldsymbol{c}}_i^\top \Phi(u)$, where $\Phi(u)$ is defined as in (14.1).

14.4 Asymptotic Properties

We introduce the matrix

$$
\boldsymbol{D}_i(\boldsymbol{F}) = \frac{1}{p}\boldsymbol{\Phi}^\top \boldsymbol{M}_{\boldsymbol{F}}\boldsymbol{\Phi} - \frac{1}{p}\boldsymbol{\Phi}^\top \boldsymbol{M}_{\boldsymbol{F}}\boldsymbol{\Phi}\boldsymbol{\lambda}_{\boldsymbol{i}}^\top \left(\frac{\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}}{n}\right)^\top \boldsymbol{\lambda}_i
$$

First, we state the assumptions made.

Assumption (1) For some constant M, $\sup \frac{1}{\sqrt{p}}\mathbb{E}\|\phi_k(u)\| \leq M, \quad k = 1, \ldots, K$.

We have also for each $i \in 1, \ldots, n, \quad \inf \boldsymbol{D}_i(\boldsymbol{F}) > 0$ □

The above assumption is on the norm of the set of basis functions. This gives a bound to the vector $\boldsymbol{\phi}_k \in \mathbb{R}^p$ that contains the discrete points on the basis function. We also require the matrix $\boldsymbol{D}_i$ to be positive definite.

Assumption (2) For some constant M,

1. $\mathbb{E}\|\boldsymbol{F}_j\|^4 \leq M$, $j = 1, \ldots, p$, and $\frac{1}{p}\sum_{j=1}^{p} \boldsymbol{F}_j \boldsymbol{F}_j^\top \longrightarrow^p \boldsymbol{\Sigma}_{\boldsymbol{F}} > 0$ for some $r \times r$ matrix $\boldsymbol{\Sigma}_{\boldsymbol{F}}$, as $p \longrightarrow \infty$;
2. $\mathbb{E}\|\boldsymbol{\lambda}_i\|^4 \leq M, i = 1 \ldots, n$, and $\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda} \longrightarrow^p \boldsymbol{\Sigma}_{\boldsymbol{\Lambda}} > 0$ for some $r \times r$ matrix $\boldsymbol{\Sigma}_{\boldsymbol{\Lambda}}$, as $n \longrightarrow \infty$. □

This assumption ensures the existence of r factors.

Assumption (3) The error terms $\epsilon_{ji}, j = 1, \ldots, p, i = 1, \ldots, n$ are independent in both directions with $\mathbb{E}(\epsilon_{ji}) = 0$, and $\mathrm{Var}(\epsilon_{ji}) = \sigma^2$; and $\mathbb{E}|\epsilon_{ji}|^8 \leq M$. Also, ϵ_{ji} is independent of ϕ_s, $\boldsymbol{\lambda}_t$, and $\boldsymbol{F}_s$ for all j, i, s, t. □

We require that the errors are independent in themselves and also of the functional and high-dimensional terms. This is a commonly seen assumption made to simplify the proofs.

Assumption (4) The tuning parameter satisfies $\alpha = o(1)$. □

This is conventionally assumed in ridge regression. This assures that the asymptotic bias of the estimator is zero.

Theorem 1 *Under Assumptions, as $n, p \longrightarrow \infty$, we have $\frac{1}{\sqrt{n}}\|\boldsymbol{C}^0 - \widehat{\boldsymbol{C}}\| \longrightarrow^p 0$.*

Next we can obtain the rate of convergence.

Theorem 2 *Under Assumptions, if $p/n \longrightarrow \rho > 0$,*

$$\sqrt{p}\frac{\|\boldsymbol{C}^0 - \hat{\boldsymbol{C}}\|}{\sqrt{n}} = O_p(1).$$

We study the case when the dimension p and the sample size n are comparable. We achieve rate $\sqrt{p}$ convergence considering $\frac{\|\boldsymbol{C}^0 - \hat{\boldsymbol{C}}\|}{\sqrt{n}}$ on the whole. This is expected that the rate of convergence for smoothing models depend on the number of discrete points observed on each curve.

References

[1] Bai, J.: Inferential theory for factor models of large dimensions. Econometrica **71**(1), 135–171 (2003)

[2] Cai, T.T., Yuan, M.: Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. The annals of statistics **39**(5), 2330–2355 (2011)

[3] Cuevas, A.: A partial overview of the theory of statistics with functional data. Journal of Statistical Planning and Inference **147**, 1–23 (2014)

[4] Eubank, R.L.: Nonparametric Regression and Spline Smoothing. CRC press (1999)

[5] Fan, J., Fan, Y., Lv, J.: High dimensional covariance matrix estimation using a factor model. Journal of Econometrics **147**(1), 186–197 (2008)

[6] Fan, J., Gijbels, I.: Local Polynomial Modelling and Its Applications. Chapman & Hall, London (1996)

[7] Febrero-Bande, M., Galeano, P., González-Manteiga, W.: Functional principal component regression and functional partial least-squares regression: An overview and a comparative study. International Statistical Review **85**(1), 61–83 (2017)

[8] Goia, A., Vieu, P.: An introduction to recent advances in high/infinite dimensional statistics. Journal of Multivariate Analysis **146**, 1–6 (2016)

[9] Green, P.J., Silverman, B.W.: Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. Chapman & Hall, London (1999)

[10] Lam, C., Yao, Q., Bathia, N.: Estimation of latent factors for high-dimensional time series. Biometrika **98**(4), 901–918 (2011)

[11] Ramsay, J.O., Hooker, G.: Dynamic Data Analysis: Modeling Data with Differential Equations. Springer, New York (2017)

[12] Ramsay, J.O., Silverman, B.W.: Applied Functional Data Analysis. Springer, New York (2002)

[13] Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer, New York (2005)

[14] Reiss, P.T., Goldsmith, J., Shang, H.L., Ogden, R.T.: Methods for scalar-on-function regression. International Statistical Review **85**(2), 228–249 (2017)

[15] Wahba, G.: Spline models for observational data, vol. 59. Siam (1990)

[16] Wand, M.P., Jones, C.M.: Kernel Smoothing. Chapman & Hall (1995)

[17] Wang, J.L., Chiou, J.M., Müller, H.G.: Functional data analysis. Annual Review of Statistics and Its Application **3**, 257–295 (2016)

[18] Yao, W., Li, R.: New local estimation procedure for a non-parametric regression function for longitudinal data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **75**(1), 123–138 (2013)

[19] Zhang, X., Wang, J.L.: From sparse to dense functional data and beyond. The Annals of Statistics **44**(5), 2281–2321 (2016)

Chapter 15
Goodness-of-fit Tests for Functional Linear Models Based on Integrated Projections

Eduardo García-Portugués, Javier Álvarez-Liébana,
Gonzalo Álvarez-Pérez and Wenceslao González-Manteiga

Abstract Functional linear models are one of the most fundamental tools to assess the relation between two random variables of a functional or scalar nature. This contribution proposes a goodness-of-fit test for the functional linear model with functional response that neatly adapts to functional/scalar responses/predictors. In particular, the new goodness-of-fit test extends a previous proposal for scalar response. The test statistic is based on a convenient regularized estimator, is easy to compute, and is calibrated through an efficient bootstrap resampling. A graphical diagnostic tool, useful to visualize the deviations from the model, is introduced and illustrated with a novel data application. The R package `goffda` implements the proposed methods and allows for the reproducibility of the data application.

15.1 Functional Linear Models

15.1.1 Formulation

Given two separable Hilbert spaces $\mathbb{H}_1$ and $\mathbb{H}_2$, we consider the regression setting with centered $\mathbb{H}_2$-valued response $\mathcal{Y}$ and centered $\mathbb{H}_1$-valued predictor $\mathcal{X}$:

Eduardo García-Portugués (✉)
Department of Statistics and UC3M-Santander Big Data Institute, Carlos III University of Madrid, Avda. Universidad 30, 28911 Leganés, Spain, e-mail: edgarcia@est-econ.uc3m.es

Javier Álvarez-Liébana
Department of Statistics and Operations Research and Mathematics Didactics, University of Oviedo, C/ Federico García Lorca, 18, 33007 Oviedo, Spain, e-mail: alvarezljavier@uniovi.es

Gonzalo Álvarez-Pérez
Department of Physics, University of Oviedo, C/ Federico García Lorca, 18, 33007 Oviedo, Spain, e-mail: gonzaloalvarez@uniovi.es

Wenceslao González-Manteiga
Department of Statistics, Mathematical Analysis and Optimization, University of Santiago de Compostela, Rúa Lope Gómez de Marzoa s/n, 15782 Santiago de Compostela, Spain, e-mail: wenceslao.gonzalez@usc.es

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_15

$$\mathcal{Y} = m(\mathcal{X}) + \mathcal{E}, \tag{15.1}$$

where $m : x \in \mathbb{H}_1 \mapsto \mathbb{E}\left[\mathcal{Y}|\mathcal{X} = x\right] \in \mathbb{H}_2$ is the regression operator and the $\mathbb{H}_2$-valued error $\mathcal{E}$ is such that $\mathbb{E}\left[\mathcal{E}|\mathcal{X}\right] = 0$. When $\mathbb{H}_1 = L^2\left([a, b]\right)$ and $\mathbb{H}_2 = L^2\left([c, d]\right)$, the Functional Linear Model with Functional Response (FLMFR; see, e.g., [15, Chapter 16]) is the most well-known parametric instance of (15.1). If the regression operator is assumed to be Hilbert–Schmidt, m is parametrizable as

$$m_\beta(\mathcal{X}) = \int_a^b \beta(s, \cdot)\mathcal{X}(s)\,\mathrm{d}s =: \langle\langle \beta, \mathcal{X}\rangle\rangle, \tag{15.2}$$

for $\beta \in \mathbb{H}_1 \otimes \mathbb{H}_2 = L^2\left([a, b] \times [c, d]\right)$ a square-integrable kernel. The present work considers this framework and is concerned with the goodness-of-fit of the family of $\mathbb{H}_2$-valued and $\mathbb{H}_1$-conditioned linear models

$$\mathcal{L} := \left\{\langle\langle \beta, \cdot\rangle\rangle : \beta \in \mathbb{H}_1 \otimes \mathbb{H}_2\right\}. \tag{15.3}$$

Any $\mathcal{X} \in \mathbb{H}_1$ and $\mathcal{Y}, \mathcal{E} \in \mathbb{H}_2$ can be represented in terms of orthonormal bases $\{\Psi_j\}_{j=1}^\infty$ and $\{\Phi_k\}_{k=1}^\infty$ as $\mathcal{X} = \sum_{j=1}^\infty x_j\Psi_j$, $\mathcal{Y} = \sum_{k=1}^\infty y_k\Phi_k$, and $\mathcal{E} = \sum_{k=1}^\infty e_k\Phi_k$, where $x_j = \langle \mathcal{X}, \Psi_j\rangle_{\mathbb{H}_1}$, $y_k = \langle \mathcal{Y}, \Phi_k\rangle_{\mathbb{H}_2}$, and $e_k = \langle \mathcal{E}, \Phi_k\rangle_{\mathbb{H}_2}$, $\forall j, k \geq 1$. Also, $\beta \in \mathbb{H}_1 \otimes \mathbb{H}_2$ can be expressed as

$$\beta = \sum_{j=1}^\infty \sum_{k=1}^\infty b_{jk}(\Psi_j \otimes \Phi_k), \quad b_{jk} = \left\langle \beta, \Psi_j \otimes \Phi_k\right\rangle_{\mathbb{H}_1 \otimes \mathbb{H}_2}, \quad \forall j, k \geq 1.$$

Therefore, the population version of the FLMFR based on (15.2) can be expressed as

$$y_k = \sum_{j=1}^\infty b_{jk}x_j + e_k, \; k \geq 1. \tag{15.4}$$

15.1.2 Model Estimation

The projection of (15.4) into the truncated bases $\{\Psi_j\}_{j=1}^p$ and $\{\Phi_k\}_{k=1}^q$ opens the way for the estimation of β given a centered sample $\{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^n$. Indeed, the truncated sample version of (15.4) is expressed as

$$\mathbf{Y}_q = \mathbf{X}_p\mathbf{B}_{p,q} + \mathbf{E}_q, \tag{15.5}$$

where $\mathbf{Y}_q$ and $\mathbf{E}_q$ are $n \times q$ matrices with the respective coefficients of $\{\mathcal{Y}_i\}_{i=1}^n$ and $\{\mathcal{E}_i\}_{i=1}^n$ on $\{\Phi_k\}_{k=1}^q$, $\mathbf{X}_p$ is the $n \times p$ matrix of coefficients of $\{\mathcal{X}_i\}_{i=1}^n$ on $\{\Psi_j\}_{j=1}^p$, and $\mathbf{B}_{p,q}$ is the $p \times q$ matrix of coefficients of β on $\{\Psi_j \otimes \Phi_k\}_{j,k=1}^{p,q}$.

Several estimators for β have been proposed; see, e.g., [16, 13, 5, 1, 14]. A popular estimation paradigm is Functional Principal Components Regression (FPCR; [15]), which considers the (empirical) Functional Principal Components (FPC) $\{\hat{\Psi}_j\}_{j=1}^p$

and $\{\hat{\Phi}_k\}_{k=1}^q$ as a plug-in for $\{\Psi_j\}_{j=1}^p$ and $\{\Phi_k\}_{k=1}^q$ underneath (15.5). Estimation by FPCR yields $\hat{\mathbf{B}}_{p,q} = \arg\min_{\mathbf{B}_{p,q}} \left\|\mathbf{Y}_q - \mathbf{X}_p\mathbf{B}_{p,q}\right\|^2 = \left(\mathbf{X}_p'\mathbf{X}_p\right)^{-1}\mathbf{X}_p'\mathbf{Y}_q$, with $j = 1,\ldots,p$ and $k = 1,\ldots,q$. The estimator $\hat{\mathbf{B}}_{p,q}$ depends on (p,q) and an automatic data-driven selection of (p,q) is of most practical interest. However, cross-validatory procedures are computationally expensive, especially since two tuning parameters must be optimized. A simple alternative for selecting q is to guarantee a certain proportion of explained variance (say, 0.99) for $\{\mathcal{Y}_i\}_{i=1}^n$. The more critical selection of p can be done by first ensuring a certain proportion of explained variance (say, 0.99) and then performing a LASSO-regularized FPCR regression (FPCR-L1 henceforth):

$$\hat{\mathbf{B}}_{p,q}^{(\lambda)} = \arg\min_{\mathbf{B}_{p,q}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left\| (\mathbf{Y}_q)_i - (\mathbf{X}_p\mathbf{B}_{p,q})_i \right\|^2 + \lambda \sum_{j=1}^{p} \left\| (\mathbf{B}_{p,q})_j \right\| \right\},$$

where the notation $(\mathbf{A})_i$ stands for the i-th row of the matrix $\mathbf{A}$. This regularization applies a row-wise penalty that enables variable selection for a given λ, which can be efficiently selected by cross-validation and its *one standard error* variant [9].

However, FPCR-L1 lacks an explicit expression for the hat matrix (in contrast with FPCR), an important handicap for the bootstrap algorithm outlined in Section 15.2.3. To combine the flexible variable selection of FPCR-L1 with the analytical form of FPCR, we propose the FPCR-L1S estimator, which firstly implements FPCR-L1 for variable selection and then performs FPCR on the selected predictors. It returns the hat matrix $\mathbf{H}_{\mathrm{C}}^{(\lambda)} = \tilde{\mathbf{X}}_{\tilde{p}}\left(\tilde{\mathbf{X}}_{\tilde{p}}'\tilde{\mathbf{X}}_{\tilde{p}}\right)^{-1}\tilde{\mathbf{X}}_{\tilde{p}}'$, where $\tilde{\mathbf{X}}_{\tilde{p}}$ is the matrix of the coefficients of the $\tilde{p}$ LASSO-selected predictors (not necessarily sorted).

Simulations [11, Section 2.4] report that FPCR-L1S outperforms FPCR.

15.2 Proposed Goodness-of-fit Tests

15.2.1 Test Statistic Genesis

Our aim is to test whether the regression operator belongs to the class of linear operators described in (15.3), that is, to test

$$\mathcal{H}_0 : m \in \mathcal{L} \quad \text{vs.} \quad \mathcal{H}_1 : m \notin \mathcal{L}.$$

To do so, we use the following lemma to characterize $\mathcal{H}_0$ in terms of the one-dimensional projections of $\mathcal{Y}$ and $\mathcal{X}$. The lemma requires from analogues of the Euclidean $(p-1)$-sphere $\mathbb{S}^{p-1} := \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x}\| = 1\}$: the $(p-1)$-sphere of $\mathbb{H}_1$ for $\{\Psi_j\}_{j=1}^\infty$, $\mathbb{S}^{p-1}_{\mathbb{H}_1,\{\Psi_j\}_{j=1}^\infty} := \{\sum_{j=1}^p x_j\Psi_j \in \mathbb{H}_1 : \|\mathbf{x}\| = 1\}$ and, analogously, $\mathbb{S}^{q-1}_{\mathbb{H}_2,\{\Phi_k\}_{k=1}^\infty}$.

Lemma 1 ($\mathcal{H}_0$ characterization on finite-dimensional directions; [11]) *Let $\mathcal{X}$ and $\mathcal{Y}$ be $\mathbb{H}_1$- and $\mathbb{H}_2$-valued random variables, respectively, $\beta \in \mathbb{H}_1 \otimes \mathbb{H}_2$, and let $\{\Psi_j\}_{j=1}^\infty$ and $\{\Phi_k\}_{k=1}^\infty$ be bases of $\mathbb{H}_1$ and $\mathbb{H}_2$, respectively. Then, the next statements are equivalent:*

i. $\mathcal{H}_0$ *holds, that is,* $m(\mathcal{X}) = \langle\langle \mathcal{X}, \beta \rangle\rangle$, $\forall \mathcal{X} \in \mathbb{H}_1$.

ii. $\mathbb{E}\Big[\big\langle \mathcal{Y} - \langle\langle \mathcal{X}, \beta \rangle\rangle, \gamma_{\mathcal{Y}}^{(q)} \big\rangle_{\mathbb{H}_2} \mathbb{1}_{\left\{\left\langle \mathcal{X}, \gamma_{\mathcal{X}}^{(p)} \right\rangle_{\mathbb{H}_1} \leq u\right\}}\Big] = 0$*, for almost every* $u \in \mathbb{R}$*,* $\forall \gamma_{\mathcal{X}}^{(p)} \in \mathbb{S}^{p-1}_{\mathbb{H}_1, \{\Psi_j\}_{j=1}^{\infty}}$, $\forall \gamma_{\mathcal{Y}}^{(q)} \in \mathbb{S}^{q-1}_{\mathbb{H}_2, \{\Phi_k\}_{k=1}^{\infty}}$*, and for all* $p, q \geq 1$.

The reader is referred to [11] for the proof of the lemma.

We use the above characterization to detect deviations from $\mathcal{H}_0$. We do so by means of the (p, q)-truncated empirical version of the doubly-projected integrated regression function in statement *ii*, that is, the residual marked empirical process

$$R_{n,p,q}\big(u, \gamma_{\mathcal{X}}^{(p)}, \gamma_{\mathcal{Y}}^{(q)}\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big\langle \hat{\mathcal{E}}_i^{(q)}, \gamma_{\mathcal{Y}}^{(q)} \big\rangle_{\mathbb{H}_2} \mathbb{1}_{\left\{\left\langle \mathcal{X}_i^{(p)}, \gamma_{\mathcal{X}}^{(p)} \right\rangle_{\mathbb{H}_1} \leq u\right\}}, \quad u \in \mathbb{R}, \tag{15.6}$$

with residual marks $\big\langle \hat{\mathcal{E}}_i^{(q)}, \gamma_{\mathcal{Y}}^{(q)} \big\rangle_{\mathbb{H}_2} = \hat{\mathbf{e}}_{i,q}' \mathbf{h}_q$ and jumps $\big\langle \mathcal{X}_i^{(p)}, \gamma_{\mathcal{X}}^{(p)} \big\rangle_{\mathbb{H}_1} = \mathbf{x}_{i,p}' \mathbf{g}_p$, where $\hat{\mathbf{e}}_{i,q}'$ represents the i-th row of the $n \times q$ matrix of residual coefficients $\hat{\mathbf{E}}_q$ on $\{\Phi_k\}_{k=1}^{q}$, $\mathbf{x}_{i,p}$ are the first p coefficients of $\mathcal{X}_i$ on $\{\Psi_j\}_{j=1}^{p}$, and $\mathbf{g}_p \in \mathbb{S}^{p-1}$ and $\mathbf{h}_q \in \mathbb{S}^{q-1}$ are the coefficients of $\gamma_{\mathcal{X}}^{(p)}$ and $\gamma_{\mathcal{Y}}^{(q)}$, respectively.

To measure the proximity of (15.6) to zero (and hence to $\mathcal{H}_0$), and following the ideas of [7] and [12], we consider a Cramér–von Mises norm on $\Pi^{(p,q)} = \mathbb{S}^{q-1}_{\mathbb{H}_2, \{\Phi_k\}_{k=1}^{\infty}} \times \mathbb{S}^{p-1}_{\mathbb{H}_1, \{\Psi_j\}_{j=1}^{\infty}} \times \mathbb{R}$, yielding the so-called Projected Cramér–von Mises (PCvM) statistic:

$$\mathrm{PCvM}_{n,p,q} = \int_{\mathbb{S}^{q-1} \times \mathbb{S}^{p-1} \times \mathbb{R}} \left[R_{n,p,q}\left(u, \mathbf{g}_p, \mathbf{h}_q\right)\right]^2 F_{n,\mathbf{g}_p}(\mathrm{d}u)\, \mathrm{d}\mathbf{g}_p\, \mathrm{d}\mathbf{h}_q,$$

where $F_{n,\mathbf{g}_p}$ is the empirical cumulative distribution function of $\{\mathbf{x}_{i,p}' \mathbf{g}_p\}_{i=1}^{n}$.

From the developments in [11], we get an easily computable form of the statistic:

$$\mathrm{PCvM}_{n,p,q} = \frac{1}{n^2} \frac{2\pi^{p/2+q/2-1}}{q\Gamma(p/2)\Gamma(q/2)} \mathrm{Tr}\left[\hat{\mathbf{E}}_q' \mathbf{A}_\bullet \hat{\mathbf{E}}_q\right], \tag{15.7}$$

where $\mathrm{Tr}(\cdot)$ denotes the trace operator and $\mathbf{A}_\bullet$ is a certain $n \times n$ symmetric matrix that only depends on $\{\mathbf{x}_{i,p}\}_{i=1}^{p}$.

15.2.2 Statistic Interpretation and Particular Cases

The statistic (15.7) can be regarded as a weighted quadratic norm:

$$\mathrm{PCvM}_{n,p,q} = \frac{1}{n^2} \frac{2\pi^{p/2+q/2-1}}{q\Gamma(p/2)\Gamma(q/2)} \sum_{k=1}^{q} \left\|(\hat{e}_{1,k}, \ldots, \hat{e}_{n,k})\right\|_{\mathbf{A}_\bullet},$$

where $\hat{\mathcal{E}}_i^{(q)} = \sum_{k=1}^{q} \hat{e}_{i,k} \Phi_k$, $i = 1, \ldots, n$, and $\|\mathbf{v}\|_{\mathbf{A}_\bullet} := (\mathbf{v}' \mathbf{A}_\bullet \mathbf{v})^{1/2}$ is a norm in $\mathbb{R}^n$ induced by $\mathbf{A}_\bullet$. Therefore, the statistic aggregates across the dimensions of the truncated response the $\mathbf{A}_\bullet$-weighted norms of the coefficients of the functional errors on $\{\Phi_k\}_{k=1}^q$. The basis of such interpretation is the next lemma (proof given in [11]).

Lemma 2 ([11]) *Assume that the functional sample* $\{\mathcal{X}_i\}_{i=1}^n$ *has pairwise* distinct *coefficients* $\{\mathbf{x}_{i,p}\}_{i=1}^n$ *on an arbitrary p-truncated basis* $\{\Psi_j\}_{j=1}^p$ *of* $\mathbb{H}_1$*. Then, for any sample size* $n \geq 1$*, the* $n \times n$ *matrix* $\mathbf{A}_\bullet$ *is positive definite.*

The general framework of the FLMFR seamless adapts to scalar response or predictor. So do the estimation methods discussed in Section 15.1.2 and the statistic (15.7). Indeed, in the case of scalar response (see, e.g., [2] and [4]), $\mathbb{H}_2 = \mathbb{R}$ is identifiable with the subspace of $L^2([c,d])$ of constant functions with basis $\{(d-c)^{-1/2}\}$ and $\beta(\cdot,\star) \equiv \beta(\cdot) \in L^2([a,b])$ is a univariate function. The statistic $\mathrm{PCvM}_{n,p,1}$ precisely corresponds to the PCvM statistic for the functional linear model with scalar response given in [12]. In the case of scalar predictor (see [3]), $\beta(\cdot,\star) \equiv \beta(\star) \in L^2([c,d])$ and $\mathrm{PCvM}_{n,1,q}$ results in a test statistic specific for such model.

15.2.3 Bootstrap Calibration and Graphical Tool

The calibration of the statistic (15.7) is done through a wild bootstrap on the residuals. We sketch next the main steps of such resampling, referring to Algorithm 1 in [11] for the specifics and its adaptation to the β-specified case.

1. Compute the statistic $\mathrm{PCvM}_{n,\tilde{p},q}$ from the residuals $\hat{\mathbf{e}}_{i,q} = \mathbf{Y}_{i,q} - \mathbf{X}_{i,\tilde{p}}\hat{\mathbf{B}}_{\tilde{p},q}^{(\lambda),\mathrm{C}}$, $i = 1, \ldots, n$, associated to the FPCR-L1S estimate $\hat{\mathbf{B}}_{\tilde{p},q}^{(\lambda),\mathrm{C}}$ (which selects $\tilde{p}$).
2. For $b = 1, \ldots, B$:
   a. Perturb the residuals as $\mathbf{e}_{i,q}^{*b} := V_i^{*b}\hat{\mathbf{e}}_{i,q}$, $i = 1, \ldots, n$, where $\{V_i^{*b}\}_{i=1}^n$ are independent zero-mean and unit-variance random variables.
   b. Using $\{\mathbf{e}_{i,q}^{*b}\}_{i=1}^n$, simulate $\{\mathbf{Y}_{i,q}^{*b}\}_{i=1}^n$ from the multivariate linear model.
   c. Fit the multivariate model from $\{(\mathbf{X}_{i,\tilde{p}}, \mathbf{Y}_{i,q}^{*b})\}_{i=1}^n$ and obtain $\hat{\mathbf{B}}_{\tilde{p},q}^{*b}$.
   d. Compute the bootstrapped statistic $\mathrm{PCvM}_{n,\tilde{p},q}^{*b}$ from the bootstrap residuals $\hat{\mathbf{e}}_{i,q}^{*b} := \mathbf{Y}_{i,q}^{*b} - \mathbf{X}_{i,\tilde{p}}\hat{\mathbf{B}}_{\tilde{p},q}^{*b}$, $i = 1, \ldots, n$.
3. Estimate the p-value by Monte Carlo as $\#\{\mathrm{PCvM}_{n,\tilde{p},q} \leq \mathrm{PCvM}_{n,\tilde{p},q}^{*b}\}/B$.

The bootstrap procedure yields as a by-product a graphical diagnostic tool of the goodness-of-fit of the FLMFR that helps visualizing the possible deviations from $\mathcal{H}_0$. The tool compares the empirical process on which the PCvM statistic is applied,

$$R_{n,p,q}\left(u, \mathbf{g}_p, \mathbf{h}_q\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\mathbf{e}}_{i,q}' \mathbf{h}_q \mathbb{1}_{\left\{\mathbf{x}_{i,p}' \mathbf{g}_p \leq u\right\}},$$

with G samples of its bootstrapped version:

$$R^{*b}_{n,p,q}\left(u,\mathbf{g}_p,\mathbf{h}_q\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{\mathbf{e}}^{*b}_{i,q})'\mathbf{h}_q\mathbb{1}_{\left\{\mathbf{x}'_{i,p}\mathbf{g}_p\leq u\right\}},\quad b=1,\ldots,G.$$

The graphical tool employs the FPC bases $\{\hat{\Psi}_j\}_{j=1}^{p}$ and $\{\hat{\Phi}_k\}_{k=1}^{q}$ and considers $\mathbf{g}_p$ and $\mathbf{h}_q$ as the canonical vectors in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. This allows to visualize the deviations from $\mathcal{H}_0$ when "it is projected" in the first FPC of $\{\mathcal{X}_i\}_{i=1}^{n}$ and the first FPC of $\{\mathcal{Y}_i\}_{i=1}^{n}$ (or any other combination thereof). Figure 15.2 shows and explains two outputs of this diagnostic tool, for the situations in which $\mathcal{H}_0$ is and is not rejected.

15.3 Application: AEMET Temperatures Dataset

The `aemet_temp` dataset in the `goffda` [10] package contains daily temperatures of $n = 73$ weather stations from the Meteorological State Agency of Spain (AEMET) during the time span 1974–2013. The dataset is split in two 20-year periods, 1974–1993 and 1994–2013, and the daily temperatures on each weather station are averaged for both periods. This results in two functional samples for the average temperatures across Spain on 1974–1993 (predictor $\mathcal{X}$) and 1994–2013 (response $\mathcal{Y}$). Both samples were smoothed with local linear estimators using cross-validated bandwidths to ease visualization. Figure 15.1 (left) shows the samples of $\mathcal{X}$ and $\mathcal{Y}$.

The PCvM test based on $\tilde{p} = 4$ (selected by FPCR-L1S with λ chosen by one standard error cross-validation) and $q = 3$ (selected such that the proportion of explained variance is 0.99) yielded a p-value equal to 0.4155 using $B = 10^4$ bootstrap replicates. Therefore, the FLMFR is not rejected. The estimated β, shown in Figure 15.1 (right), reveals a temperature increment on the latter period with respect to the former, a conclusion supported by the predominance of positive values on the $\hat{\beta}$ surface and the positiveness of almost all the temperature curves. The diagnostic tool in Figure 15.2 (left) shows no remarkable deviations of the residual marked empirical process from $\mathcal{H}_0$. The PCvM test rejects emphatically the simple hypotheses $\mathcal{H}_0 : \beta = 0$ and $\mathcal{H}_0 : \beta(s,t) = \mathbb{1}_{\{s=t\}}$ (stationary-temperature hypothesis; right panel in Figure 15.2), thus corroborating a significant change in the temperatures between both periods. The diagnostic tool for the latter hypothesis reveals that the non-stationarity is due to the relations between the second FPC of $\{\mathcal{X}_i\}_{i=1}^{n}$ and $\{\mathcal{Y}_i\}_{i=1}^{n}$, both related with the variation shape of the temperature curves along the year.

15.4 Software: `goffda` R Package

The R package `goffda` [10] implements all the methods described and allows for replication of the data application. The implementation of the critical parts of the goodness-of-fit tests, such as the computation of the $\mathbf{A}_\bullet$ matrix and the computation of the PCvM statistic, are implemented in C++ (through `Rcpp` [6]) for the sake of efficiency. The `goffda` package relies on the `fdata` class from the `fda.usc` [8] package, so it is fully compatible with the latter.

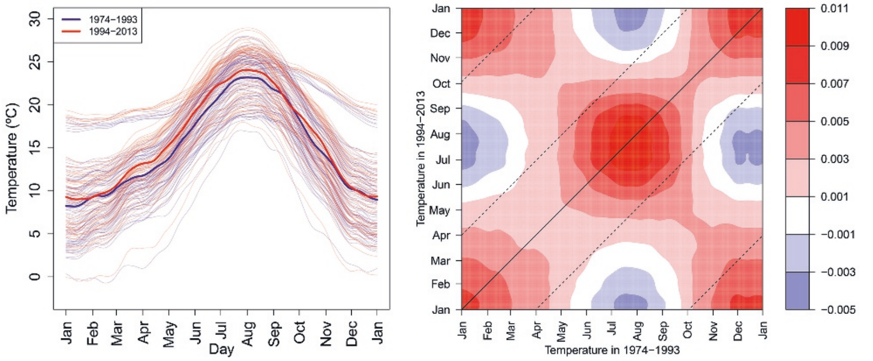


Fig. 15.1 Left: Temperatures of 73 AEMET weather stations for the periods 1974–1983 ($\mathcal{X}$) and 1994–2013 ($\mathcal{Y}$), along with their means. Right: FPCR-L1S estimator $\hat{\beta}$ for the FLMFR.

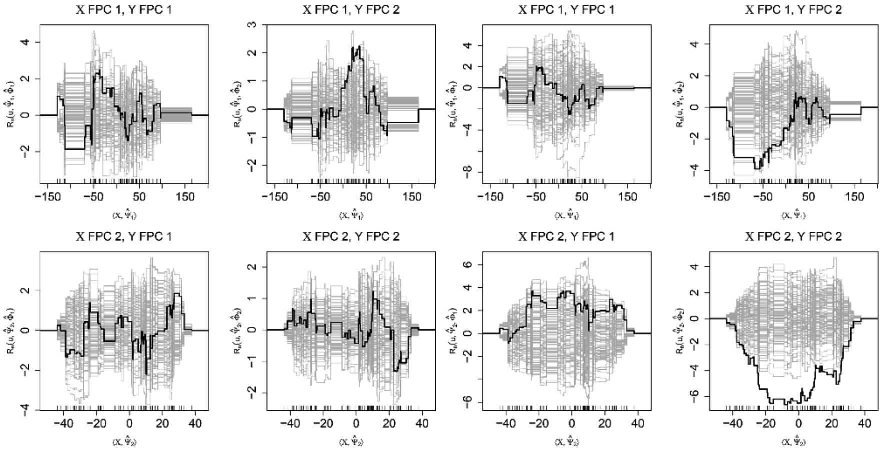


Fig. 15.2 Graphical tool of the PCvM test. The black curve represents the observed process $R_{n,p,q}\left(u, \mathbf{e}_j, \mathbf{e}_k\right)$ for its projections on the j-th FPC of $\{\mathcal{X}_i\}_{i=1}^n$ and the k-th FPC of $\{\mathcal{Y}_i\}_{i=1}^n$, $j, k = 1, 2$. The grey curves stand for the bootstrapped processes under $\mathcal{H}_0$, i.e., $R_{n,p,q}^{*b}\left(u, \mathbf{e}_j, \mathbf{e}_k\right)$, $b = 1, \ldots, 100$. The left 2×2 panel shows the diagnostic output for $\mathcal{H}_0 : m \in \mathcal{L}$ in the AEMET temperatures dataset. The non-rejection of $\mathcal{H}_0$ is manifested in the centrality of the observed process within the bootstrapped ones. The right 2×2 panel shows the diagnostic for $\mathcal{H}_0 : \beta(s, t) = \mathbb{1}_{\{s=t\}}$, with rejection of $\mathcal{H}_0$ evidenced by the outlyingness of $R_{n,p,q}\left(u, \mathbf{e}_2, \mathbf{e}_2\right)$.

The main functions of `goffda` are: `flm_est` (several estimation methods for the FLMFR); `Adot` (efficient implementation of the $\mathbf{A}_\bullet$ matrix); `flm_stat` (computation of (15.7)); `flm_test` (implementation of the test with its bootstrap resampling). `flm_est` and `flm_test` deal seamlessly with either functional/scalar responses/predictors.

Acknowledgements The authors gratefully thank Prof. Manuel Febrero-Bande for discussions and for providing access to the dataset of raw AEMET temperatures. The first author acknowledges support from grants PGC2018-097284-B-I00 and IJCI-2017-32005 from the Spanish Ministry of Economy and Competitiveness (co-funded with FEDER funds). The second author acknowledges support from grant PGC2018-099549-B-I00 from the same agency. The first and fourth authors acknowledge support from grant MTM2016-76969-P also from the same agency. The authors gratefully acknowledge the computing resources of the Supercomputing Center of Galicia (CESGA).

References

[1] Benatia, D., Carrasco, M., Florens, J.P.: Functional linear regression with functional response. J. Econometrics **201**(2), 269–291 (2017)

[2] Cardot, H., Ferraty, F., Sarda, P.: Functional linear model. Statist. Prob. Lett. **45**(1), 11–22 (1999)

[3] Chiou, J.M., Müller, H.G., Wang, J.L., Carey, J.R.: A functional multiplicative effects model for longitudinal data, with application to reproductive histories of female medflies. Statist. Sinica **13**(4), 1119–1133 (2003)

[4] Crambes, C., Kneip, A., Sarda, P.: Smoothing splines estimators for functional linear regression. Ann. Statist. **37**(1), 35–72 (2009)

[5] Crambes, C., Mas, A.: Asymptotics of prediction in functional linear regression with functional outputs. Bernoulli **19**(5B), 2627–2651 (2013)

[6] Eddelbuettel, D., François, R.: Rcpp: Seamless R and C++ integration. J. Stat. Softw. **40**(8), 1–18 (2011)

[7] Escanciano, J.C.: A consistent diagnostic test for regression models using projections. Econometric Theory **22**(6), 1030–1051 (2006)

[8] Febrero-Bande, M., Oviedo de la Fuente, M.: Statistical computing in functional data analysis: The R package fda.usc. J. Stat. Softw. **51**(4), 1–28 (2012)

[9] Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**(1), 1–22 (2010)

[10] García-Portugués, E., Álvarez-Liébana, J.: goffda: Goodness-of-fit tests for functional data. R package version 0.0.6 (2019). https://CRAN.R-project.org/package=goffda.

[11] García-Portugués, E., Álvarez-Liébana, J., Álvarez-Pérez, G., González-Manteiga, W.: A goodness-of-fit test for the functional linear model with functional response. arXiv:1909.07686 (2019)

[12] García-Portugués, E., González-Manteiga, W., Febrero-Bande, M.: A goodness-of-fit test for the functional linear model with scalar response. J. Comp. Graph. Stat. **23**(3), 761–778 (2014)

[13] He, G., Müller, H.G., Wang, J.L., Yang, W.: Functional linear regression via canonical analysis. Bernoulli **16**(3), 705–729 (2010)

[14] Imaizumi, M., Kato, K.: PCA–based estimation for functional linear regression with functional responses. J. Multivariate Anal. **163**, 15–36 (2018)

[15] Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer Series in Statistics. Springer, New York (2005)

[16] Yao, F., Müller, H.G., Wang, J.L.: Functional linear regression analysis for longitudinal data. Ann. Statist. **33**(6), 2873–2903 (2005)

Chapter 16
From High-dimensional to Functional Data: Stringing Via Manifold Learning

Harold A. Hernández-Roig, M. Carmen Aguilera-Morillo and Rosa E. Lillo

Abstract The study of high-dimensional data is becoming a common trend in modern research. Recently, stringing emerged as a methodology to treat high-dimensional sample vectors as realizations of smooth stochastic processes. Under the hypothesis of noisy and order-perturbed measurements, stringing introduces smooth transitions between predictors and takes advantage of Functional Data Analysis (FDA) to study the data. Once a functional representation is achieved, it is possible to visualize intrinsic patterns, or fit functional regression models. We propose manifold learning as an alternative to multidimensional scaling in the reordering step. In a simulation study we show that our proposal achieves smaller relative order errors, and that it can recover more complex relationships between predictors.

16.1 Introduction

High-dimensional data refer to scenarios in which the dimension p, the number of features or predictor variables, is so large that calculations become extremely difficult. Often, in high-dimensional data the number of observations n is much smaller than p ($n \ll p$). In these scenarios, modeling is a challenging problem that has been addressed under strong assumptions, such as sparsity constraints [5]. The prevailing approach for $n \ll p$ problems is usually related to dimensionality reduction, and variable selection, in other words, excluding features from the analysis. On the other hand, data visualization techniques—such as parallel coordinates—fail

Harold A. Hernández-Roig (✉)
Universidad Carlos III de Madrid and uc3m-Santander Big Data Institute, Spain,
e-mail: haroldantonio.hernandez@uc3m.es

M. Carmen Aguilera-Morillo
Universitat Politècnica de València and uc3m-Santander Big Data Institute, Spain,
e-mail: mdagumor@eio.upv.es

Rosa E. Lillo
Universidad Carlos III de Madrid and uc3m-Santander Big Data Institute, Spain,
e-mail: rosaelvira.lillo@uc3m.es

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_16

in these cases, due to the high-dimensionality of p. A different view, mentioned in [3], is that of Andrew's Plots [1], in which the high dimensional vectors are expanded in terms of trigonometric functions. Nevertheless, none of these techniques consider the intrinsic relationship between predictors.

Stringing was introduced in [3] as a methodology to map high-dimensional vectors to the infinite-dimensional function space. Once transformed, all tools from FDA are available to study high-dimensional data. This methodology had early beginnings in classification of gene expression profiles [11], and it has been extended to other functional regression models, such as Cox's [3]. In these scenarios, typical $n \ll p$, it is possible to take advantage of the high-dimensionality of the data to transform them into functions. Then, functional modeling can be applied to estimate different types of responses. Moreover, this methodology provides a sophisticated graphical representation of multivariate data that reveals its characteristics and intrinsic patterns.

Briefly, stringing assumes that data is observed with an unknown and randomly permuted order of the predictors. The methodology consists of reordering these predictors to achieve smooth transitions between the components of each data vector. Finally, it considers the transformed vectors as realizations of a smooth stochastic process. In the original paper [3], the authors apply Unidimensional Scaling (UDS) to the columns of the design matrix, obtaining positions in $\mathbb{R}$ for each predictor. UDS is the unidimensional version of Multidimensional Scaling (MDS), thus, it finds a one-dimensional representation that preserves distances between predictors in the higher-dimensional space. Functional representation is achieved with Functional Principal Components Analysis (FPCA), or any other smoothing technique (see [7] for examples).

Two basic problems arise from UDS-stringing: (i) complex relationships between predictors, say non-linearity, could be invisible for some distances; and (ii) while reordering, we might be losing important features of the data that cannot be represented in $\mathbb{R}$, but in higher dimensions (say $\mathbb{R}^2$, or $\mathbb{R}^3$). This work focuses on the first problem. We propose stringing via manifold learning, an alternative—based on Isometric Feature Mapping (Isomap) [9] and Locally Linear Embedding (LLE) [8] algorithms—that is able to recover more complex relationships between predictors, and assigns positions in the real line for each of them. In a simulation study we show that the reordering error decreases when manifold learning replaces UDS.

16.2 Stringing via Manifold Learning

The original approach of stringing consists of projecting a collection $\mathbf{x}_1, \ldots, \mathbf{x}_p$ of predictors from $\mathbb{R}^n$ to $\hat{s}_1, \ldots, \hat{s}_p$ in $\mathbb{R}^l$, being $l \ll n$. Only the case $l = 1$ has been addressed in the literature. UDS is used to estimate the best ranks d^*_{rs} of pairwise distances that could lead to the lower-dimensional representation $\hat{s}_1, \ldots, \hat{s}_p \in \mathbb{R}$. This intermediate step has similarities with seriation or sequencing [6]. The difference is that stringing takes advantage of the inherent order of $\mathbb{R}$, and treats the reordered components of each subject as realizations of a smooth stochastic process.

Assume we observe p features (or predictors) for n subjects, and that we can arrange data in a matrix $\mathbf{X}_{n\times p}$ with rows:

$$\left\{x_{i\cdot} = \left(x_{i1}, \ldots, x_{ip}\right)\, ;\; x_{i\cdot}^{\tau} \in \mathbb{R}^{p}\right\}_{i=1}^{n} .$$

Each row is generated by a hidden smooth stochastic process $\{Z_i(s), s \in I \subset \mathbb{R}\}$, with support points $s_j \in I$, $j = 1, \ldots, p$, so that $Z_i(s_j) = x_{ij}$ is an entry of $\mathbf{X}$. The stringing methodology is developed under the hypothesis that what we actually observe is a matrix $\tilde{\mathbf{X}}$ with a randomly permuted order of the columns. In such cases, treating the rows as realizations of a smooth function is meaningless and a previous reordering step is needed.

UDS is perhaps the simplest tool to recover the best unidimensional configuration of the columns $\tilde{\mathbf{x}}_j$ of $\tilde{\mathbf{X}}$. Through any suitable pairwise-column distance (dissimilarities are also possible) we can build a matrix $D = (d_{rs})_{1\le r,s\le p}$ with entries:

$$d_{rs} = d(\tilde{\mathbf{x}}_r, \tilde{\mathbf{x}}_s);\quad \tilde{\mathbf{x}}_r, \tilde{\mathbf{x}}_s \in \mathbb{R}^n.$$

The procedure finds minimizing dissimilarities d^*_{rs}, in the target space $\mathbb{R}$, of the *stress*:

$$S^2\left(\hat{s}\right) = \min_{\{d^*_{rs}:\, d^*_{rs}\sim d_{rs}\}} \frac{\sum_{r<s}(d^*_{rs} - \hat{d}_{rs})^2}{\sum_{r<s} \hat{d}^2_{rs}},$$

where $d^*_{rs} \sim d_{rs}$ means *monotonically related* quantities ($d_{rs} < d_{uv} \implies d^*_{rs} \le d^*_{uv},\ \forall\ r < s,\ u < v$); and the $\hat{d}_{rs}$ represent point-to-point distances of a configuration $\hat{s} \subset \mathbb{R}$. Details can be consulted in [6].

Therefore, with UDS we can assign a support point $\hat{s}_j \in \mathbb{R}$ to each predictor indexed by $j = 1, \ldots, p$. In [3], the new order of the predictors is characterized by a permutation ψ_p, called the *stringing function*, such that $\hat{s}_{\psi_p(1)} < \hat{s}_{\psi_p(2)} < \ldots < \hat{s}_{\psi_p(p)}$. For each predictor j with rank order $\psi_p(j)$, and for a fix T, we could also define its regularized position $s_{jp} = \dfrac{j-1}{p-1}\cdot T$. The purpose is to normalize the resulting domain to $[0, T]$.

As an alternative to UDS, we assume that the predictors $\tilde{\mathbf{x}}_j = (x_{1j}, \ldots, x_{nj})^{\tau} \in \mathbb{R}^n$ (columns of $\tilde{\mathbf{X}}$) are the result of mapping the coordinates $\{s_j \in \mathcal{M}\}$ of an underlying l-dimensional smooth manifold $\mathcal{M}$. Thus, through manifold learning it is possible to recover the unknown coordinates of these vectors in a lower, but more complex, l-dimensional space. In this paper we focus on the case $l = 1$, analogously to UDS-stringing. In this context we propose Isomap [9] and LLE [8] algorithms to compute the estimates $\{\hat{s}_j \in \mathcal{M}\}$.

The motivation behind this approach can be illustrated with the *Swiss-roll example*, Fig. 16.1. Consider the rectangle $\tilde{\mathcal{M}}$ covering $[0, 10]\times[0, 10] \subset \mathbb{R}^2$, Fig. 16.1.a. We roll this two-dimensional manifold and map it to $\mathbb{R}^3$ using the transformation $(x, y) \rightarrow (x\cos(x), y, x\sin(x))$, Fig. 16.1.c. MDS preserves Euclidean distances across dimensions. Thus, if we apply this method to our data in $\mathbb{R}^3$, points like A and B will be positioned closer than they really are in the underlying rectangle.

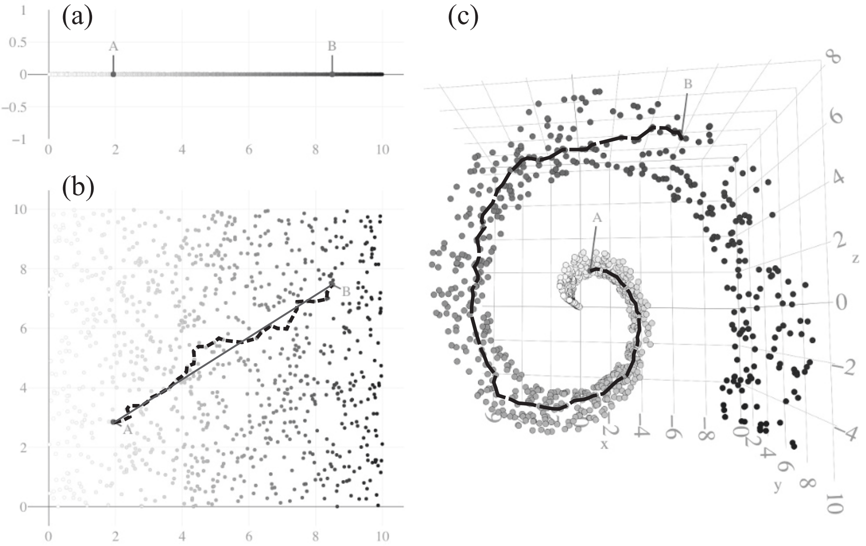


Fig. 16.1 The *Swiss-roll example*: a two-dimensional manifold $\hat{\mathcal{M}}$ (rectangle in $[0, 10] \times [0, 10] \subset \mathbb{R}^2$) is mapped to $\mathbb{R}^3$. **(a)** The one dimensional configuration to be estimated with stringing. **(b)** The true distance between points A and B in $\hat{\mathcal{M}}$ (*bold line*) is approximated by the shortest path (*dashed line*): the result of joining neighboring points. **(c)** The coordinates in $\mathbb{R}^3$ exhibit a Swiss-roll shape. The points A and B seem to be closer in the higher-dimensional space, according to the Euclidean distance. The *dashed line* approximates the geodesic distance in $\hat{\mathcal{M}}$, by connecting neighboring points in $\mathbb{R}^3$.

The Isomap algorithm approximates the geodesic distances $\{d_{rs}^{\mathcal{M}}\}$—those distances in the underlying manifold—by means of the shortest paths between any of the pairs $\tilde{\mathbf{x}}_r, \tilde{\mathbf{x}}_s \in \mathbb{R}^n$. These paths are the result of joining neighboring points, defined as the κ-nearest according to the Euclidean distance in $\mathbb{R}^n$. Finally, it applies MDS to estimate the $\{\hat{s}_j\}$, using as inputs the approximations $\{\hat{d}_{rs}^{\mathcal{M}}\}$. This is a global approach to manifold learning, as it preserves the global geometry of $\mathcal{M}$.

In the Swiss-roll example, Isomap estimates the shortest paths $\{\hat{d}_{AB}^{\tilde{\mathcal{M}}}\}$ (*dashed line* in Fig. 16.1.c) between A and B, by joining neighboring points in $\mathbb{R}^3$. This estimation approximates the true Euclidean distance in $\mathbb{R}^2$ (*bold line* in Fig. 16.1.b) where the manifold $\tilde{\mathcal{M}}$ lies. Moreover, if we fix the target dimension to be $l = 1$, manifold learning still retrieves a fair projection onto the x-axis (Fig. 16.1.a). To overcome the problem of fixing l, we adopt the Local Quality criterion (Q_{local}) [4], and estimate the optimum number of neighbors $\kappa_{\max} < p$ that improves the estimation of the coordinates in $\mathbb{R}$. Increasing the number $\kappa_{\max}$ makes the outcome of Isomap closer to that from classical MDS, so we expect Isomap-stringing to work at least as well as the UDS-stringing.

The second algorithm, LLE, preserves local neighborhood information in the manifold, without estimating the true geodesic distances. It fixes a suitable number κ and reconstructs each point $\hat{\tilde{\mathbf{x}}}_i = \sum_{j=1}^{\kappa} \hat{w}_{ij}\tilde{\mathbf{x}}_j$, in terms of its κ-nearest neighbors

$N_i^{\kappa} = \{\tilde{\mathbf{x}}_j\}_{j=1}^{\kappa}$, and some optimal weights $\hat{w}_{ij}$—in the sense that they minimize the *reconstruction error* $\sum_{i=1}^{p} \|\hat{\tilde{\mathbf{x}}}_i - \tilde{\mathbf{x}}_i\|^2$; subject to $\sum_j \hat{w}_{ij} = 1$, and $\hat{w}_{ij} = 0, \forall x_j \notin N_i^{\kappa}$. The coordinates $\{s_i \in \mathcal{M}\}_{i=1}^{p}$, best reconstructed by the weights $\{\hat{w}_{ij}\}_{j=1}^{\kappa}$, are estimated by minimizing the *embedding cost function*:

$$\sum_{i=1}^{p} \|s_i - \sum_{j=1}^{\kappa} \hat{w}_{ij} s_j\|^2.$$

Under some constraints that make the objective function invariant under translation, rotation, and change in scale, the problem is reduced to the estimation of the bottom $l+1$ eigenvectors of the sparse $p \times p$ matrix $M = \left(I_p - \hat{W}\right)^{\tau} \left(I_p - \hat{W}\right)$. The "bottom" eigenvectors refer to those with the $l+1$ smallest eigenvalues, I_p is the identity matrix of size $p \times p$, and $\hat{W}$ is the matrix of optimal weights $(\hat{w}_{ij})_{1 \le i,j \le p}$. Details regarding the LLE algorithm can be consulted in the original paper [8]. Once more, we focus on the target dimension $l = 1$, and adopt the Q_{local} criterion to compute the optimum κ_{max}.

Finally, once a configuration $\hat{s} \subset \mathbb{R}$ is achieved, we can compute the regularized nodes $\{s_{jp} \in [0,T] \subset \mathbb{R}\}$, and represent the underlying smooth stochastic process that generates the data by means of the Karhunen-Loève (K-L) expansion [7]. The model can be expressed as follows:

$$\begin{aligned} x_{ij} &= Z_i(s_{jp}) + \epsilon_{ij} \\ &= \mu(s_{jp}) + \sum_{k=1}^{\infty} \xi_{ik} \phi_k(s_{jp}) + \epsilon_{ij}, \end{aligned}$$

where μ is the mean function of the underlying stochastic process, and $\{\phi_k\}$ is a sequence of orthonormal eigenfunctions, in the $L^2\left([0,T]\right)$ sense, of the covariance operator $A_G : L^2([0,T]) \longrightarrow L^2([0,T])$. This operator is defined as:

$$(A_G \cdot f)(t) = \int_{[0,T]} G(s,t) f(s) ds,$$

for any $f \in L^2\left([0,T]\right)$. The kernel $G(s,t) = \mathbb{E}\left[(X(t) - \mu(t)) \cdot (X(s) - \mu(s))\right]$, for any $s,t \in [0,T]$; is the covariance function of the stochastic process X. The principal component scores $\{\xi_k\}$ are zero-mean uncorrelated random variables satisfying $\xi_k = \int_{[0,T]} \left(X(t) - \mu(t)\right) \phi_k(t) dt$. The error terms are all independent and identically distributed (i.i.d.) $\epsilon_{ij} \sim N(0, \sigma^2)$.

16.3 Simulation Study

We simulate data from noisy Ornstein-Uhlenbeck (O-U) processes, and study the performance of stringing via manifold learning. This is the case of zero-mean stochastic processes $\{U_t : t \in [0,T]\}$, characterized by the covariance function:

$$G(s,t) = P\exp\left(\alpha|t-s|\right).$$

We generate the O-U processes by means of the truncated K-L expansion:

$$U(t) = \sum_{k=1}^{K} \xi_k \phi_k(t); \qquad t \in [0,T] \subset \mathbb{R}.$$

Following [10], it is possible to estimate the eigenvalues $\{\lambda_k\}$ for each O-U process using the formula: $\lambda_k = \dfrac{2P\alpha}{\alpha^2 + b_k^2}$, where the numbers b_k are the positive solutions of:

$$\tan\left(b_k \frac{T}{2}\right) = \frac{\alpha}{b_k} \qquad (k \text{ is odd}); \qquad \tan\left(b_k \frac{T}{2}\right) = -\frac{b_k}{\alpha} \qquad (k \text{ is even}).$$

The eigenfunctions $\{\phi_k(t)\}$, normalized in $[0,T]$ are:

$$\frac{\cos\left(b_k\left(t-\frac{T}{2}\right)\right)}{\left[\frac{T}{2}\left(1+\dfrac{\sin(b_kT)}{b_kT}\right)\right]^{1/2}} \qquad (k \text{ is odd}); \qquad \frac{\sin\left(b_k\left(t-\frac{T}{2}\right)\right)}{\left[\frac{T}{2}\left(1-\dfrac{\sin(b_kT)}{b_kT}\right)\right]^{1/2}} \qquad (k \text{ is even}).$$

For i.i.d. $\xi_k \sim N(0,1)$ and i.i.d. $z_{ij} \sim N(0,1)$, we generate noisy O-U realizations:

$$x_{ij} = \sum_{k=1}^{K} \xi_k \phi_k(t_j) + \sigma \cdot z_{ij},$$

where the error variance is computed in terms of the signal-to-noise ratio (SNR):

$$SNR = \frac{\frac{1}{n\cdot p}\sum_{i,j}|u_{ij}|}{\sigma} \in \{\infty, 6.81, 2.68\}.$$

We study different n/p ratios: $n = 30, p = 100$; $n = p = 50$; and $n = 60, p = 100$. The nodes $\{t_j\}$ are p regularized positions on the interval $[0,4]$. For fixed $P = 1$, $\alpha = 0.1$ we take $K = 14$ basis functions and scores.

We apply stringing to 400 matrices $\tilde{\mathbf{X}}_{n\times p}$ generated from noisy O-U processes, but with randomly-permuted columns. We compare Isomap-stringing and LLE-stringing with UDS-stringing based on Euclidean distance and Pearson Correlation. To assess the performance of each method, we use the relative order error (ROE) introduced in [3]:

$$ROE = \frac{\sum_{j=1}^{p}|o_j^S - o_j|}{\mathbb{E}\left(\sum_{j=1}^{p}|o_j^R - o_j|\right)} = \frac{\sum_{j=1}^{p}|o_j^S - o_j|}{\dfrac{(p-1)(p+1)}{3}},$$

where o_j denotes the true rank for each predictor indexed by $j = 1,\ldots,p$; o_j^R the rank of predictor j after the random permutation; and o_j^S the rank induced by stringing.

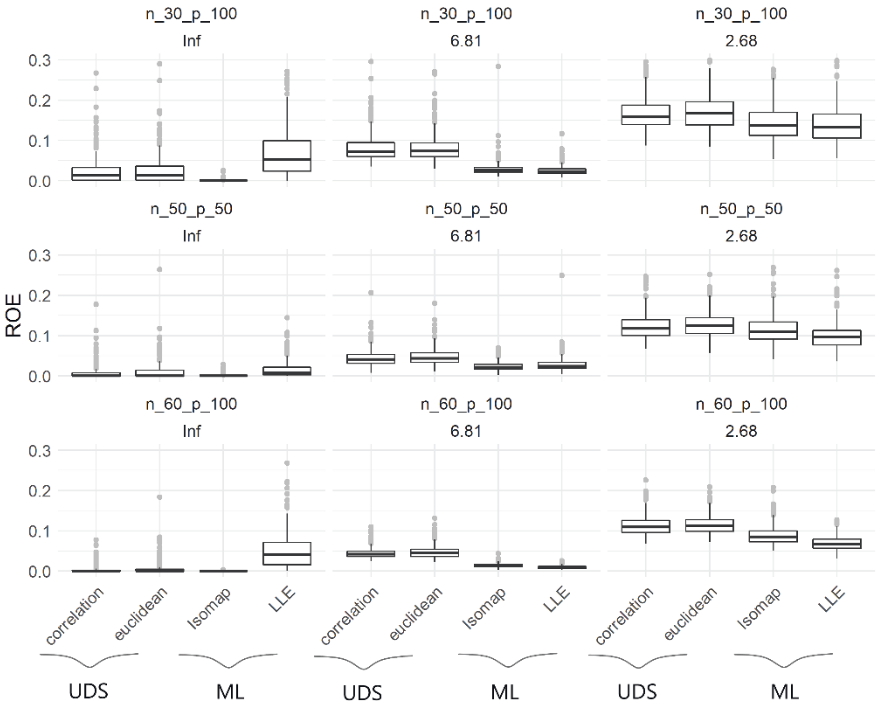


Fig. 16.2 Boxplots of ROE values computed on 400 simulated O-U processes. Columns represent different SNR: $\{\infty, 6.81, 2.68\}$. Rows represent three different n/p ratios: $\{30/100; 50/50; 60/100\}$. Stringing is carried out via Unidimensional Scaling (UDS) and Manifold Learning (ML).

Fig. 16.2 includes the resulting boxplots of ROE values, under different SNR (and therefore different variances). Comparisons with Isomap-stringing and LLE-stringing show that manifold learning outperforms UDS: lower medians, smaller interquartile range (IQR) and fewer outliers. In this study, Isomap seems to be the best algorithm when $SNR \to \infty$ ($\sigma \to 0$). On the other hand, LLE performs better in noisier scenarios, no matter the ratio n/p. Isomap and LLE are the most consistent algorithms under the hypothesis of measurement errors.

16.4 Discussion

The results of this study indicate that stringing via manifold learning achieves a better reordering of the data, compared to the previous version of the methodology. If predictors are allowed to be in a manifold, then it is possible to estimate smoother transitions between them. This hypothesis is consistent with the intrinsic complexity of high-dimensional data vectors. Thus, combining our proposal with FDA can result in an effective representation of general high-dimensional data.

We hypothesize that our proposal, combined with functional regression, could lead to new insights into the problem of high-dimensional data modeling. This idea has been explored for UDS-stringing in [3], as an alternative to the sparsity constraints from penalized regression. Also, the essence of manifold learning: reducing dimension, could be a key factor to extend stringing. The idea of estimating positions in $\mathbb{R}^2$ or $\mathbb{R}^3$ for each predictor—instead of assigning ranks—was mentioned in [3], but has not been addressed in the literature.

The applicability to real data, such as genetic expression arrays [11, 3], makes stringing and its generalization an attractive research topic. Recently, some extensions to more complex datasets have been published. In [2] stringing is applied to study functional connectivity in patients with Alzheimer. In [12] stringing is included as part of a functional test for high-dimensional covariance matrix, with application to mitochondrial calcium concentration.

References

[1] Andrews, D.F.: Plots of High-Dimensional Data. Biometrics **28**(1), 125–136 (1972)

[2] Chen, C.J., Wang, J.L.: A New Approach for Functional Connectivity via Alignment of Blood Oxygen Level-Dependent Signals. Brain Connectivity **9**(6), 464–474 (2019)

[3] Chen, K., Müller, H.-G., Wang, J.: Stringing High-Dimensional Data for Functional Analysis. Journal of the American Statistical Association **106**(493), 275–284 (2011)

[4] Chen, L., Buja, A.: Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. Journal of the American Statistical Association **104**(485), 209–219 (2009)

[5] Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. Chapman and Hall/CRC (2015)

[6] Mardia, K., Kent, J., Bibby, J.: Multivariate Analysis. Academic Press (1979)

[7] Ramsay, J., Silverman, B.: Functional Data Analysis (2nd ed). Springer Science+Business Media, Inc (2005)

[8] Roweis, S.T., Saul, L.K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science **290**(5500), 2323–2326 (2000)

[9] Tenenbaum, J.B., de Silva, V., Langford, J.C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science **290**(5500), 2319–2323 (2000)

[10] Trees, H.L.V., Bell, K.L., Tian,Z.: Detection Estimation and Modulation Theory, Part I: Detection, Estimation, and Filtering Theory (2nd ed). Wiley (2013)

[11] Wu, P.S., Müller, H.G.: Functional embedding for the classification of gene expression profiles. Bioinformatics **26**(4), 509–517 (2010)

[12] Zhang, T., Wang, Z., Wan, Y.: Functional test for high-dimensional covariance matrix, with application to mitochondrial calcium concentration. Statistical Papers (2019)

Chapter 17
Functional Two-sample Tests Based on Empirical Characteristic Functionals

Zdeněk Hlávka and Daniel Hlubinka

Abstract Two-sample tests for functional data based on empirical characteristic functionals are proposed. The test statistic is of Cramér–von Mises type with integration over a preselected family of probability measures, say Q, leading a computationaly feasible and powerful test statistic. The choice of the probability measure Q is discussed and the empirical size and power of the resulting two-sample functional tests are investigated in a small simulation study.

17.1 Introduction

Functional data analysis already became a standard [11, 6, 7, 9] with many tools obtained as a generalization of a corresponding multivariate method. In this contribution, we investigate the general functional two-sample problem and propose a new two-sample functional test statistic based on empirical characteristic functionals.

Assuming two functional random samples, say $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, the problem of testing the null hypothesis of equality of the respective mean functions, i.e.,

$$H_0 : m_X(.) = m_Y(.)$$

has already been extensively investigated, see [3] for an overview. Slightly different hypothesis is studied in [8], namely

$$H_0 : \forall_t \; X(t) =_{\mathcal{L}} Y(t)$$

Zdeněk Hlávka
Univerzita Karlova, Dept. of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Sokolovská 83, Praha 8, Czechia, e-mail: hlavka@karlin.mff.cuni.cz

Daniel Hlubinka (✉)
Univerzita Karlova, Dept. of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Sokolovská 83, Praha 8, Czechia, e-mail: hlubinka@karlin.mff.cuni.cz

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_17

testing simultaneously the distribution of all projections where $=_{\mathcal{L}}$ denotes the equality in distribution.

In the following, instead of comparing only the mean functions or testing the distributions of projections, we are interested in testing a more general null hypothesis of equality of entire functional distributions:

$$\mathcal{H}_0 : \phi_X = \phi_Y \tag{17.1}$$

where ϕ_X and ϕ_Y denotes, respectively, the *characteristic functional* (CF) of the X and Y sample. The definition and properties of CF and *empirical CF* (ECF) are summarized in Section 17.2. A two-sample test statistic based on a distance between two ECFs is proposed in Section 17.3. Finally, a small simulation study in Section 17.4 investigates small sample properties of the ECF-based two-sample test.

17.2 Empirical Characteristic Functional

In what follows, we consider functional random variables with values in the space of continuous functions or in the space of measurable square integrable functions, i.e., $X : \Omega \to C[0,1]$ or $X : \Omega \to \mathcal{L}_2[0,1]$, where the domain is as usually (and wlog) chosen to be $[0,1]$.

The CF of X is $\phi_X(u) = \mathrm{E}\exp(i\langle u, X\rangle)$ for $u \in C^*[0,1]$ or $u \in \mathcal{L}_2^*[0,1]$, the dual space of $C[0,1]$ or $\mathcal{L}_2[0,1]$, respectively. Due to the properties of CF, it is sufficient to consider just $u \in \mathcal{L}_2^*[0,1] = \mathcal{L}_2[0,1]$ for both options in which case $\langle u, X\rangle = \int_0^1 u(t)X(t)\mathrm{d}t$.

The ECF of a functional random sample $X_1, \dots, X_n$ is

$$\widetilde{\phi}_X(u) = \frac{1}{n}\sum_{k=1}^{n} \exp\bigl(i\langle u, X_k\rangle\bigr).$$

The functional data are not observed continuously in most cases. We may consider all X_i's to be observed on a regular grid of points $t_j = j/N$, $j = 0, 1, \dots, N$ since the generalisation to different observation points is straightforward. The ECF is then

$$\widehat{\phi}_X(u) = \frac{1}{n}\sum_{k=1}^{n} \exp\bigl(i\langle u, X_k\rangle_d\bigr),$$

where $\langle u, X\rangle_d = \sum_{i=1}^{N} u(t_i)X(t_i)(t_i - t_{i-1}) = \frac{1}{N}\sum_{i=1}^{N} u(t_i)X(t_i)$.

17.3 Cramér–von Mises Type of Statistics

Our Cramér–von Mises two-sample test is based on two random samples $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$ with the test statistic

$$\int_{\mathcal{L}_2[0,1]} |\widehat{\phi}_X(u) - \widehat{\phi}_Y(u)|^2 \mathrm{d}Q(u), \tag{17.2}$$

where Q is some probability measure on the space $\mathcal{L}_2[0,1]$ discussed later. The squared distance $|\widehat{\phi}_X(u) - \widehat{\phi}_Y(u)|^2$ of the ECFs may be rewritten as

$$\left(\frac{1}{n}\sum_{k=1}^{n}\cos\langle u, X_k\rangle_d - \frac{1}{m}\sum_{\ell=1}^{m}\cos\langle u, Y_\ell\rangle_d\right)^2 + \left(\frac{1}{n}\sum_{k=1}^{n}\sin\langle u, X_k\rangle_d - \frac{1}{m}\sum_{\ell=1}^{m}\sin\langle u, Y_\ell\rangle_d\right)^2 \tag{17.3}$$

and we obtain after some calculation and using the trigonometric identity the final form

$$\frac{1}{n^2}\sum_{k,j=1}^{n}\cos\langle u, X_k - X_j\rangle_d + \frac{1}{m^2}\sum_{\ell,j=1}^{m}\cos\langle u, Y_\ell - Y_j\rangle_d - \frac{2}{mn}\sum_{k=1}^{n}\sum_{\ell=1}^{m}\cos\langle u, X_k - Y_\ell\rangle_d. \tag{17.4}$$

17.3.1 Choice of Q

The measure Q is some probability measure on the space $\mathcal{L}_2[0,1]$. We propose to use a special form of this measure, namely some special form of a Gaussian measure. Hence, we consider a random function U with all finite-dimensional distribution being multivariate normal distribution. Since the data are observed on a discrete grid of N points, it is sufficient to consider a random vector $U_N = \big(U(t_1), \ldots, U(t_N)\big)$ following zero mean N-dimensional normal distribution with the variance matrix $\boldsymbol{V} = (v_{i,j})_{i,j=1}^{N}$. For a fixed discretely observed function x, we have

$$\langle U, x\rangle_d = \frac{1}{N}\sum_{i=1}^{N} U(t_i)x(t_i) \sim \mathcal{N}\left(0, \frac{1}{N^2}\sum_{j,k=1}^{N} x(t_i)x(t_j)v_{i,j}\right) = \mathcal{N}\big(0, \sigma^2(x)\big),$$

where $\sigma^2(x) = \frac{1}{N^2}x^T\boldsymbol{V}x$ and $\mathrm{E}_Q \cos\langle U, x\rangle_d$ becomes

$$\mathrm{E}_Q \cos\left(\frac{1}{N}\sum_{i=1}^{N} U(t_i)x(t_i)\right) = \exp\left(-\frac{1}{2}\sigma^2(x)\right) \tag{17.5}$$

and the test statistic (17.2) based on the two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ becomes

$$\begin{aligned} T = \frac{1}{n^2} \sum_{k,j=1}^{n} \exp\left(-\frac{1}{2N^2}(X_k - X_j)^T \boldsymbol{V}(X_k - X_j)\right) \\ + \frac{1}{m^2} \sum_{k,j=1}^{m} \exp\left(-\frac{1}{2N^2}(Y_k - Y_j)^T \boldsymbol{V}(Y_k - Y_j)\right) \\ - \frac{2}{nm} \sum_{k=1}^{n} \sum_{j=1}^{m} \exp\left(-\frac{1}{2N^2}(X_k - Y_j)^T \boldsymbol{V}(X_k - Y_j)\right). \end{aligned} \qquad (17.6)$$

The null hypothesis will be rejected for large values of the test statistic T, i.e., for

$$T \geq c(\alpha), \qquad (17.7)$$

where $c(\alpha)$ denotes critical value such that $P(T \geq c(\alpha)|\mathcal{H}_0) = \alpha$. In the following, the critical value will be approximated by the permutation principle [2].

17.3.2 The Matrix $\boldsymbol{V}$

The performance of the test largely depends on the matrix $\boldsymbol{V}$ introduced in Section 17.3.1. We propose several possibilities and our test is then compared with other two-sample tests in a small simulation study.

The most simple choice is to set $\boldsymbol{V} = \mathcal{I}_N$ but the following possibilities should have better power.

Variance matrix of a Gaussian process: This proposal follows classical "random projection" approach. It is considered that U is a Gaussian process, usually a Wiener process and $\boldsymbol{V} = \Sigma_W$ is the variance matrix of the process observed at $j/N, j = 1, 2, \ldots, N$.

The observations: We consider $n + m$ iid $Z_\ell \sim \mathcal{N}(0, 1)$, and

$$U = \frac{1}{\sqrt{n+m}} \left[\sum_{j=1}^{n} Z_j X_j + \sum_{k=1}^{m} Z_{k+n} Y_k \right].$$

Then

$$z_{q,r} = \frac{1}{N^2(n+m)} \left[\sum_{j=1}^{n} X_j(t_q) X_j(t_r) + \sum_{k=1}^{m} Y_k(t_q) Y_k(t_r) \right]$$

and we can set $\boldsymbol{V} = \boldsymbol{Z} = (z_{q,r})_{q,r=1,\ldots,N}$.

Sample covariance matrix: By centering the (functional) observations, we actually obtain $\boldsymbol{V} = \hat{\Sigma}$, where $\hat{\Sigma}$ denotes the sample variance matrix of the observed N-dimensional random vectors (approximating the functional observations).

Notice that the quadratic forms in exponential functions in (17.6) look similarly to Hotelling's T^2 test statistic, where the matrix $\boldsymbol{V}$ is chosen as the inverse of the sample covariance matrix. Therefore, further possible choices of the matrix $\boldsymbol{V}$ could be the inverse of the matrix $\hat{\Sigma}$ or Σ_W. Note that the inverse of $\hat{\Sigma}$ generally does not exist but, depending on the number of observations, first d eigenvectors and eigenvalues can be used to calculate a simple approximation. Interestingly, the eigenvectors (or eigenfunctions) tend to recover the direction (in the functional space) that separates the two sets of functional observations, see also the discussion in [5].

Eigenvectors and eigenvalues: Denote by $\lambda_1 \geq \lambda_2 \geq \dots$ the ordered eigenvalues, and by $e_1, e_2, \dots$ the corresponding orthogonal eigenfunctions of the covariance operator of the combined dataset. Consider $d \geq 1$ and iid random variables $Z_\ell \sim \mathcal{N}(0,1)$, and define

$$U = \sum_{\ell=1}^{d} \frac{1}{\sqrt{\lambda_\ell}} e_\ell Z_\ell .$$

The theoretical eigenfunctions and eigenvalues are replaced by eigenvectors $\hat{e}_\ell$ and eigenvalues $\hat{\lambda}_\ell$ of the empirical variance matrix $\hat{\Sigma}$ in practical applications. Then for some $1 \leq d \leq \min(m+n, N)$ define

$$e_{q,r} = \sum_{\ell=1}^{d} \frac{1}{\hat{\lambda}_\ell} \hat{e}_\ell(t_q) \hat{e}_\ell(t_r).$$

In the following, we denote the resulting matrix $\boldsymbol{V} = (e)_{q,r=1,\dots,N} = \hat{\Sigma}_d^{-1}$. The choice of d is discussed later.

17.4 Simulation and Comparison

We start by investigating the empirical power against the 'location shift' alternative. Following [4, Section 5], we generate two functional samples

$$X_i(t) = \mu_x(t) + \varepsilon_{x,i}(t) \tag{17.8}$$

and

$$Y_i(t) = \mu_y(t) + \varepsilon_{y,i}(t),$$

where the mean functions are $\mu_x(t) = (1, 2.3, 3.4, 1.5)(1, t, t^2, t^3)^\top$ and $\mu_y(t) = \mu_x(t) + 2\delta(1,2,3,4)(1,t,t^2,t^3)^\top/\sqrt{30}$ so that the parameter δ controls the difference between $\mu_x(t)$ and $\mu_y(t)$. The subject-effect functions $\varepsilon_{\cdot,i}(t)$ are defined as a random linear combination of 11 orthonormal basis vectors $\psi_w(t)$ (such that $\psi_1(t) = 1$, $\psi_{2\omega}(t) = \sqrt{2}\sin(2\pi\omega t)$, $\psi_{2\omega+1}(t) = \sqrt{2}\cos(2\pi\omega t)$, for $\omega = 1, \dots, 5$) with coefficients $b_{\cdot,i,w} \sim N(0, 1.5\rho^w)$, for $w = 1, \dots, 11$.

In this section, we set $\rho = 0.5$. The choice $\delta = 0$ means that the null hypothesis is satisfied and we investigate the empirical size. An example of two data sets generated under the alternative, with $\delta = 0.5$, is plotted in Figure 17.1. Note that these two samples were generated in the same way as samples 2 and 3 in [4, Section 5.2].

We compare empirical sizes and powers of the proposed two-sample ECF-based test (ECF) with various variance matrices $\boldsymbol{V}$, described in Section 17.3.2, to tests implemented in R library fdANOVA [4]. Many of these tests are based on the usual

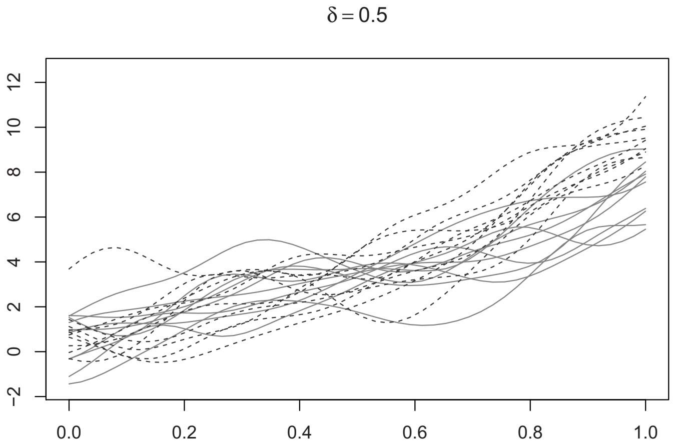


Fig. 17.1 Random samples $X_1, \ldots, X_{10}$ (solid lines) and $Y_1, \ldots, Y_{10}$ (dashed lines) generated according to the algorithm described in Section 17.4 with $\delta = 0.5$ and $\rho = 0.5$.

univariate (pointwise) F-statistics, say $F_n(t)$ for $t \in (0, 1)$, that are combined into a single test statistic:

GPF: globalizing pointwise F test, $T^{\text{GPF}} = \int F_n(t)dt$,
Fmaxb: maximizing pointwise F test, $T^{\text{Fmaxb}} = \max F_n(t)$,

Another approach is based on testing k projections of the original functions by combining p-values [1] that are based on:

ANOVA: ANOVA F-test statistic,
ATS: ANOVA-type statistic,
WTPS: Wald type permutation statistic.

Similarly to the choice of the random process (and matrix $\boldsymbol{V}$) in Section 17.3.2, the projections are generated either as Gaussian white noise (G) or Brownian motion (B). A detailed description of the function `fanova.tests()` in R library `fdANOVA` is given in [4].

In the first two columns of Table 17.1, we can see that the empirical size of all tests is close to the nominal level $\alpha = 5\%$ both for $n = m = 10$ and $n = m = 20$ observations.

For $n = m = 10$, the empirical power is smallest for ECF tests with $\boldsymbol{V} = \Sigma_W^{-1}$ (7.8%) and $\boldsymbol{V} = \boldsymbol{Z}$ (27.6%). The empirical power of most tests lies between 47% and 60%. Somewhat higher power, almost 70%, has been obtained for Fmaxb and ECF tests with $\boldsymbol{V} = \hat{\Sigma}_6^{-1}$. Using $d = 8$ eigenvectors, the highest empirical power (90.5%) is observed for the ECF test with $\boldsymbol{V} = \hat{\Sigma}_8^{-1}$.

		$\delta = 0$		$\delta = 0.5$	
method	test/$\boldsymbol{V}$	$n = 10$	$n = 20$	$n = 10$	$n = 20$
GPF		6.1	6.2	58.0	88.4
FMAXB		6.1	3.6	69.8	98.0
	ANOVA	3.5	3.3	51.1	90.1
P-Gauss	ATS	5.0	4.9	54.2	89.6
	WTPS	3.8	3.6	47.8	90.5
	ANOVA	2.7	3.4	54.7	87.7
P-BM	ATS	3.4	3.5	55.5	87.0
	WTPS	2.6	3.1	47.5	88.0
	$\mathcal{I}$	4.7	4.0	53.9	89.3
	Σ_W	4.4	4.8	57.5	87.8
	Σ_W^{-1}	5.8	4.8	7.8	11.8
	$\boldsymbol{Z}$	4.8	4.9	27.6	52.1
ECF	$\hat{\Sigma}$	4.5	5.6	43.2	76.6
	$\hat{\Sigma}_2^{-1}$	4.2	5.7	47.6	79.4
	$\hat{\Sigma}_6^{-1}$	5.2	5.9	67.6	95.2
	$\hat{\Sigma}_8^{-1}$	4.2	4.9	**90.5**	**99.9**

Table 17.1 Empirical size ($\delta = 0$) and empirical power ($\delta = 0.5$) (in %) of two-sample functional tests, nominal level $\alpha = 0.05$, $\rho = 0.5$, $N = 50$ gridpoints, equally sized samples ($n = m$), 1000 simulations with 1000 permutations. Bold font denotes the highest observed empirical power.

Results for $n = m = 20$ are similar but observed differences are smaller because the power of most tests is close to 90%.

		F-statistic		ECF		
		$n = 10$	$n = 20$	$\boldsymbol{V}$	$n = 10$	$n = 20$
GPF		7.6	5.4	$\mathcal{I}$	6.3	9.2
FMAXB		6.9	6.2	Σ_W	22.4	54.1
	ANOVA	4.7	4.4	Σ_W^{-1}	8.5	14.8
P-Gauss	ATS	4.9	3.3	$\boldsymbol{Z}$	28.6	59.1
	WTPS	4.9	4.5	$\hat{\Sigma}$	**53.2**	**89.5**
	ANOVA	3.5	3.1	$\hat{\Sigma}_2^{-1}$	6.4	4.9
P-BM	ATS	3.6	3.9	$\hat{\Sigma}_6^{-1}$	9.8	7.0
	WTPS	3.1	3.0	$\hat{\Sigma}_8^{-1}$	10.0	10.6

Table 17.2 Empirical power ($\sigma_y = 2$) (in %) of two-sample functional tests, nominal level $\alpha = 0.05, \rho = 0.5$, $N = 50$ gridpoints, equally sized samples ($n = m$), 1000 simulations with 1000 permutations. Bold font denotes the highest observed empirical power.

The study of the empirical power against the 'change-of-scale' alternative is summarized in Table 17.2; the random functions $X_i(t)$ are still generated according to (17.8) while the second sample is changed to

$$Y_i^{\sigma}(t) = \mu_y(t) + \sigma_y \varepsilon_{y,i}(t),$$

with $\delta = 0$ (implying that $\mu_x(.) = \mu_y(.)$) and with additional parameter $\sigma_y > 0$ controlling the variance. As may be expected, the empirical power of the F-statistic-based tests shown in Table 17.2 is very close to the nominal test level. The best power is obtained for the ECF-based test with $\boldsymbol{V} = \hat{\Sigma}$, the sample covariance matrix. On the other hand, the ECF-based tests using $\boldsymbol{V}$ based on the inversion of (some) covariance matrix do not perform very well.

We conclude that the ECF test with the matrix $\boldsymbol{V}$ approximating the inverse covariance matrix leads to the best results against the 'location shift' alternative while the ECF test with $\boldsymbol{V} = \hat{\Sigma}$ leads to the best results against the 'change-of-scale' alternative. Interestingly, the ECF test outperforms the F-statistic-based tests implemented in library `fdANOVA` even against the 'location shift' alternative.

Acknowledgements The research was supported by Czech Science Foundation, grant GAČR 18-08888S.

References

[1] Cuesta-Albertos, J.A., Febrero-Bande, M.: A simple multiway ANOVA for functional data. Test **19**, 537–557 (2010)

[2] Good, P.: Permutation, Parametric, and Bootstrap Tests of Hypotheses. Springer-Verlag, New York (2005)

[3] Górecki, T., Smaga, Ł.: A comparison of tests for the one-way ANOVA problem for functional data. Comp. Stat. **30**, 987–1010 (2015)

[4] Górecki, T., Smaga, Ł.: fdANOVA: an R software package for analysis of variance for univariate and multivariate functional data. Comp. Stat. **34**, 571–597 (2019)

[5] Hall, P., Poskitt, D.S., Presnell, B.: A functional data-analytic approach to signal discrimination, Technometrics **43**, 1–9 (2001)

[6] Horváth, L., Kokoszka, P.: Inference for Functional Data with Applications. Springer (2012)

[7] Hsing, T., Eubank, R.: Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. Wiley (2015)

[8] Jiang, Q., Hušková, M., Meintanis, S.G., Zhu, L.: Asymptotics, finite-sample comparisons and applications for two-sample tests with functional data. Journal of Multivariate Analysis **170**, 202–220 (2019)

[9] Kokoszka, P., Reimherr, M.: Introduction to Functional Data Analysis. Chapman and Hall/CRC (2017)

[10] Meintanis, S.G.: A review of testing procedures based on the empirical characteristic function. South African Statistical Journal **50**, 1–14 (2016)

[11] Ramsay, J.O., Silverman, B.W.: Applied Functional Data Analysis: Methods and Case Studies. Springer (2007)

Chapter 18
Some Remarks on the Nelson–Siegel Model

Lajos Horváth

Abstract We discuss some results in functional data analysis where the mean and the covariance are expended in some theoretically justified basis.

18.1 Introduction and Motivation

We consider functional observations $X_1(t), X_2(t), \ldots, X_N(t)$ defined on the interval $\mathcal{T}$. It is popular to assume in financial models that

$$X_i(t) = \sum_{\ell=1}^{K} b_{i,\ell,0} f_\ell(t; \lambda_0) + \epsilon_i(t), \quad \text{with} \quad E\epsilon_i(t) = 0, \quad t \in \mathcal{T}, \quad 1 \le i \le N, \quad (18.1)$$

where the random coefficients satisfy

$$b_{i,\ell,0} = c_{\ell,0} + e_{i,\ell} \quad \text{with} \quad Ee_{i,\ell} = 0, \quad t \in \mathcal{T}, \qquad 1 \le \ell \le K \quad \text{and} \quad 1 \le i \le N. \quad (18.2)$$

Under assumptions (18.1) and (18.2)

$$EX_i(t) = \sum_{\ell=1}^{K} c_{\ell,0} f_\ell(t; \lambda_0), \quad t \in \mathcal{T} \quad \text{and} \quad 1 \le i \le N,$$

i.e. the mean of the observations can be written as a linear combination of the functions $f_1(t; \lambda_0), f_2(t; \lambda_0), \ldots, f_K(t; \lambda_0)$, where the functions $f_1, f_2, \ldots, f_K$ are known and $\lambda_0 \in R^d$ is the true value of an unknown parameter. In several applications λ_0 is fixed and set to a value based on previous experiments or the choice is justified by theoretical arguments. It is generally not estimated, so the curves $f_k(t; \lambda_0)$ are treated as given, deterministic functions. These functions are called functional factors, or simply factors. The motivation for (18.1) and (18.2) is the popular Nelson–Siegel

Lajos Horváth (✉)
University of Utah, Department of Mathematics, Salt Lake City UT 84112, U.S.A.,
e-mail: horvath@math.utah.edu

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_18

model [13] (cf. [6]) and its extensions. Further motivations and applications of the model in (18.1) and (18.2) are given in [1, 3, 4, 5, 7, 16, 17]. The parameter of the model is $\mathbf{a} = (c_1, c_2, \ldots, c_K, \boldsymbol{\lambda}^\top)^\top \in R^{K+d}$, whose true value is $\mathbf{a}_0 = (c_{1,0}, c_{2,0}, \ldots, c_{K,0}, \boldsymbol{\lambda}_0^\top)^\top$. [10] use least squares to estimate the value of $\boldsymbol{\lambda}_0$. The estimator $\hat{\boldsymbol{\lambda}}_N$ minimizes the least squares loss function

$$U_N(\mathbf{a}) = \frac{1}{N}\sum_{i=1}^{N}\int\left(X_i(t) - \sum_{\ell=1}^{K} c_\ell f_\ell(t; \boldsymbol{\lambda})\right)^2 dm(t), \tag{18.3}$$

where $m(t)$ is a suitably chosen weight function. The estimator $\hat{\mathbf{a}}_N$ is thus defined by

$$\hat{\mathbf{a}}_N = \operatorname{argmax}_{\mathbf{a}\in\mathbf{A}} U_N(\mathbf{a}).$$

[10] establish the almost sure consistency and the asymptotic normality of $\hat{\boldsymbol{\lambda}}_N$. These results are used to provide tests to check the validity of the model defined by (18.1) and (18.2). Following the methodology of [2], heteroscedastic functional errors ϵ_i in (18.1) and heteroscedastic multivariate errors $e_{i,\ell}$ in (18.2) are allowed. In contrast to [2], [10] does not assume that the times, when the second order properties of ϵ_i and/or $e_{i,\ell}$ change, are known. So the errors in the Nelson–Siegel model might not be stationary, the observation is segmented into periods of mean stationarity. [8] investigated a similar model. They established tests to find changes in the mean of the observations when the variance of the observations is a function of time.

18.2 Mathematical Interpretation of the Nelson–Siegel Model

We rewrite (18.1) and (18.2) as

$$X_i(t) = \sum_{\ell=1}^{K} c_{\ell,0} f_\ell(t; \boldsymbol{\lambda}_0) + \sum_{\ell=1}^{K} e_{i,\ell,0} f_\ell(t; \boldsymbol{\lambda}_0) + \epsilon_i(t), \quad t \in \mathcal{T}, \quad 1 \le i \le N. \tag{18.4}$$

Hence

$$EX_i(t) = \mu(t) = \sum_{\ell=1}^{K} c_{\ell,0} f_\ell(t; \boldsymbol{\lambda}_0)$$

and K functions completely determine the mean of the observations. Since the functions are linearly independent in the space of square integrable functions, we can assume without loss of generality that they are orthogonal. The errors are decomposed into two parts in (18.4), one part is spanned by $f_1, f_2, \ldots, f_K$ and the second part is $\epsilon_i(t)$ and it is assumed that these two terms are independent or at least uncorrelated. The decomposition of (18.4) is similar to the projection method which has been successfully used in functional data analysis. [9] and [11] provide detailed accounts of the applications of projections and several data examples.

Let $\phi_1, \phi_2, \ldots,$ be an orthonormal basis in $L^2[\mathcal{T}]$, the Hilbert space of square integrable functions on $\mathcal{T}$. The inner product in $L^2[\mathcal{T}]$ is defined by

$$\langle f, g\rangle = \int_{\mathcal{T}} f(t)g(t)dt$$

and the corresponding norm is $\|\cdot\|_{\mathcal{T}}$. We write X_i in the standard nonparametric form

$$X_i(t) = \mu(t) + \eta_i(t) \ \text{ with } \ E\eta_i(t) = 0, \ \ t \in \mathcal{T}, \ 1 \le i \le N. \tag{18.5}$$

The projection of $X_i(t)$ into the direction of $\phi_\ell(t)$ is

$$\langle X_i, \phi_\ell\rangle = \mu_\ell + \eta_{i,\ell}, \ \ 1 \le i \le N, 1 \le \ell < \infty,$$

where $\mu_\ell = \langle \mu, \phi_\ell\rangle$ and $\eta_{i,\ell} = \langle X_i, \phi_\ell\rangle$. Thus we obtain the Karhunen–Loéve expansion of $X_i(t)$

$$X_i(t) = \sum_{\ell=1}^{\infty} \mu_\ell \phi_\ell(t) + \sum_{\ell=1}^{\infty} \eta_{i,\ell}\phi_\ell(t),$$

which can be written as

$$X_i(t) = \sum_{\ell=1}^{K} \mu_\ell \phi_\ell(t) + \sum_{\ell=1}^{K} \eta_{i,\ell}\phi_\ell(t) + \bar{\eta}_i(t), \ \ t \in \mathcal{T}, \ 1 \le i \le N.$$

Of course, $\bar{\eta}_i(t) = 0$ if and only if $\phi_1(t), \phi_2(t), \ldots, \phi_K(t)$ span $\mu(t)$. So the Nelson–Siegel model picks K functions which completely explain the mean.

The most often used method to study random curves is the functional principle component analysis, which uses the orthonormal eigenfunctions of

$$C(t, s) = E(X_i(t) - \mu(t))(X_i(s) - \mu(s)),$$

assuming that the sample is stationary. The eigenvalues $\tau_1 \ge \tau_2 \ge \ldots$ and the corresponding eigenfunctions $\psi_1(t), \psi_2(t), \ldots$ are the solutions of the eigenvalue problem

$$\tau_j \psi_j(t) = \int_{\mathcal{T}} C(t, s)\psi_j(s)ds, \ \ 1 \le j < \infty. \tag{18.6}$$

(the norm of the eigenfunctions is 1). The principle component analysis gives the best approximation for the error term in (18.5). However, the scores $\langle \mu, \psi_\ell\rangle, 1 \le \ell \le K$ might not provide any information about $\mu(t)$ or the approximation for $\mu(t)$ in terms of the first K eigenfunctions of the covariance function $C(t, s)$ is rather poor. As usual, $C(t, s)$ unknown but it is easily approximated by the empirical covariance function

$$\hat{C}_N(t, s) = \frac{1}{N}\sum_{i=1}^{N}(X_i(t) - \hat{\mu}_N(t))(X_i(s) - \hat{\mu}_N(s)), \ \ \text{with} \ \ \hat{\mu}_N(t) = \frac{1}{N}\sum_{i=1}^{N} X_i(t).$$

The empirical eigenvalues $\hat{\tau}_1 \ge \hat{\tau}_2 \ge \ldots \ge \hat{\tau}_N \ge 0$ and eigenfunctions $\hat{\psi}_1(t), \hat{\psi}_2(t), \ldots, \hat{\psi}_N(t)$ defined by

$$\hat{\tau}_j\hat{\psi}_j(t) = \int_{\mathcal{T}} \hat{C}_N(t,s)\hat{\psi}_j(s)ds, \quad 1 \le j \le N \tag{18.7}$$

are used for the (empirical) projections. As usual, the norm of $\hat{\psi}_j$ is 1. If the $L^2[\mathcal{T}\times\mathcal{T}]$ norm of $\hat{C}_N(t,s) - C(t,s)$ goes to 0 in probability, we have immediately the consistency of the estimators $\hat{\tau}_j$. The eigenfunctions of $\psi_j(t)$ and $\hat{\psi}_j(t)$ are not unique, even in case of distinct eigenvalues, they are determined up to a sign only. [9] provides a survey of the estimation of eigenvalues and eigenfunctions. [15] and [14] contain theory and applications of estimates in case or repeated eigenvalues.

Let $0 \le \alpha \le 1$ and define

$$A_\alpha(t,s) = (1-\alpha)C(t,s) + \alpha\mu(t)\mu(s), \quad (t,s) \in \mathcal{T}\times\mathcal{T}.$$

It is clear that $A(t,s)$ is a non negative definite function, so we can define again the eigenvalues $\gamma_{1,\alpha} \ge \gamma_{2,\alpha} \ge \ldots$ and the corresponding eigenfunctions $\zeta_{1,\alpha}(t), \zeta_{2,\alpha}(t), \ldots$ as the solutions of

$$\gamma_{j,\alpha}\zeta_{j,\alpha}(t) = \int_{\mathcal{T}} A(t,s)\zeta_{j,\alpha}(s)ds, \quad 1 \le j < \infty. \tag{18.8}$$

If $\alpha = 0$, then (18.8) reduces to (18.6) and if $\alpha = 1$, then the only nonzero eigenvalue is 1 and the corresponding eigenfunction is constant times $\mu(t)$. Hence the projections into the directions of $\zeta_j(t), 1 \le j \le K$ are a compromise between capturing the mean and the covariance of stationary observations. The parameter α is chosen such that the mean squared error

$$\mathrm{MSE}(\alpha) = E\left\| X_1(t) - \sum_{\ell=1}^{K} \langle X_1, \zeta_\ell\rangle \zeta_\ell(t)\right\|_{\mathcal{T}}^2$$

is minimised. We suggest estimating α from the sample. Let

$$\hat{A}_{N,\alpha}(t,s) = (1-\alpha)\hat{C}_N(t,s) + \alpha\hat{\mu}(t)\hat{\mu}(s), \quad (t,s) \in \mathcal{T}\times\mathcal{T}$$

and $\hat{\gamma}_{1,\alpha} \ge \hat{\gamma}_{2,\alpha} \ge \ldots \ge \hat{\gamma}_{N,\alpha} \ge 0$ and $\hat{\zeta}_{1,\alpha}(t), \hat{\zeta}_{2,\alpha}(t), \ldots, \hat{\zeta}_{N,\alpha}(t)$ are the eigenvalues and the orthonormal eigenfunctions of the empirical $\hat{A}_{N,\alpha}(t,s)$. The empirical version of $\mathrm{MSE}(\alpha)$ is

$$\widehat{\mathrm{MSE}}(\alpha) = \frac{1}{N}\sum_{i=1}^{N}\left\| X_i(t) - \sum_{\ell=1}^{K} \langle X_i, \hat{\zeta}_{\ell,\alpha}\rangle \hat{\zeta}_{\ell,\alpha}(t)\right\|_{\mathcal{T}}^2$$

and we suggest using the minimiser of $\widehat{\mathrm{MSE}}(\alpha)$ in practice. We note that if $X_1, X_2, \ldots, X_N$ of (18.5) and $E\|X_1\|_{\mathcal{T}}^2 < \infty$, then by the ergodic theorem in Hilbert spaces we have that for all $0 \le \alpha \le 1$

$$\left\|\hat{A}_{N,\alpha} - A_\alpha\right\| \to 0 \text{ a.s.}$$

Hence the methods in Section 2.5 of [9] can be used to show that

$$\left|\hat{\gamma}_{\ell,\alpha} - \gamma_{\ell,\alpha}\right| \to 0 \ \text{ a.s.}$$

and

$$\left\|\hat{\zeta}_{\ell,\alpha} - s_\ell \zeta_{\ell,\alpha}\right\| \to 0 \ \text{ a.s.},$$

where $s_1, s_2, \ldots, s_K$ are random signs. Hence all statistical methods based on principal components and their empirical versions can be easily modified to employ the eigenfunctions of A_α or $\hat{A}_{N,\alpha}$.

In several applications the sum of the observations plays an important role and in this case it might be more suitable to use the long run covariance function of the stationary sequence $X_\ell, -\infty < \ell < \infty$. The long run covariance function is defined as

$$D(t,s) = \sum_{\ell=-\infty}^{\infty} \operatorname{var}\left(X_\ell(t), X_\ell(s)\right).$$

Kernel based estimators are used to estimate $D(t,s)$ from the sample and the consistency of the estimator is established under various conditions. [12] consider several estimators of the log run covariance function when the means of several functional populations are compared. They discuss the consistency of procedures, using different type long run covariance estimators, in detail. For surveys we refer again to [9] and [11]. Using the eigenfunctions of $D(t,s)$ we approximate well the random term but we cannot be sure if the mean is also captured. The method discussed above can be easily modified when $C(t,s)$ is replaced with $D(t,s)$. Thus we obtain better projections for the mean and $D(t,s)$,

References

[1] Bank for International Settlements: Zero-coupon yield curves: technical documentation. BIS Paper **25** (2005)

[2] Bardsley, P., Horváth, L., Kokoszka, P., Young, G.: Change point tests in functional factor models with application to yield curves. Econometrics Journal **20**, 373–403 (2017)

[3] Chambers, D.R., Carleton, W.T., Waldman, D.W.: A new approach to estimation of the term structure of interest rates. Journal of Financial and Quantitative Analysis **19**, 233–252 (1984)

[4] Christensen, J.H.E., Diebold, F.X., Rudebusch, G.D.: The affine arbitrage–free class of Nelson—Siegel term structure models. Journal of Econometrics **164**, 4–20 (2011)

[5] Diebold, X.F., Rudebusch, G.D., Aruoba, S.: The macroeconomy and the yield curve: a dynamic latent factor approach. Journal of Econometrics **131**, 339–338 (2006)

[6] Diebold, X.F., Rudebusch, G.D.: Yield Curve Modeling and Forecasting. Princeton University Press (2013)

[7] Filipovič, D.: A note on the Nelson–Siegel family. Mathematical Finance **9**, 349–359 (1999)

[8] Górecki, T., Horváth, L., Kokoszka, P.: Change point detection in heteroscedastic time series. Econometrics & Statistics **20**, 86–117 (2017)

[9] Horváth, L., Kokoszka, P.: Inference for Functional Data with Applications. Springer, New York (2012)

[10] Horváth, L., Kokoszka, P., VanderDoes, J., Wang, S.: Inference in Dynamic Nelson–Siegel Models. In progress (2020)

[11] Horváth, L., Rice, G.: A survey of functional data analysis and the functional analysis of variance problem. Revista Matemática Complutense **28**, 505–548 (2015)

[12] Horváth, L., Rice, G.: Testing equality of means when the observations are from functional time series. Journal of Time Series Analysis **36**, 84–108 (2015)

[13] Nelson, C.R., Siegel, A.F.: Parsimonious modeling of yield curves. The Journal of Business **60**, 473–489 (1987)

[14] Petrovich, J., Reimherr, M.: Asymptotic properties of principal component projections with repeated eigenvalues. Statistics & Probability Letters **130**, 42–48 (2017)

[15] Reimherr, M.: Functional regression with repeated eigenvalues. Statistics & Probability Letters **107**, 62–70 (2015)

[16] Svensson, L.E.O.: Estimating and Interpreting Forward Interest Rates: Sweden 1992 - 1994. NBER Working Paper No. **w4871** (1994)

[17] Yallup, P.J.: Models of the yield curve and the curvature of the implied forward rate function. Journal of Banking & Finance **36**, 121–135 (2012)

Chapter 19
Modeling the Effect of Recurrent Events on Time-to-event Processes by Means of Functional Data

Francesca Ieva, Marta Spreafico and Davide Burba

Abstract In this paper we propose a methodological framework for modeling information carried out by a longitudinal process by means of functional data, within a survival framework targeting the time-to-event process of interest. In particular, the longitudinal process is represented by the compensator of a marked point process the recurrent events are supposed to derive from. By means of Functional Principal Component Analysis (FPCA), a suitable dimensional reduction of these objects is carried out in order to plug them into a survival Cox regression model. In doing so, we enrich the information available for modeling survival with relevant dynamic features, whose time-varying nature is properly taken into account. Such methodology is applied to data provided by the healthcare division of Lombardia regional district in Italy, related to patients hospitalized for Heart Failure (HF) between 2000 and 2012, who assume multiple drugs over time. The model enables personalized predictions, quantifying the effect of personal behaviors and therapeutic patterns on long-term survival.

19.1 Introduction

A recurrent event is an event which may occur more than once. Many situations in clinical practice can be modeled in the framework of recurrent events, for instance the process of re-hospitalizations of chronic patients over time, drug purchases and many others. In this paper we look at the recurrent events for a set of individuals as particular stochastic processes, namely *marked point processes for recurrent events*. We assume the form of the generating Cox model for counting processes [1] for

Francesca Ieva (✉)
MOX, Department of Mathematics, Politecnico di Milano, P.zza Leonardo da Vinci 32, 20133 Milano (IT) & CADS - Center for Analysis, Decision and Society, Human Technopole, Milano, Italy, e-mail: francesca.ieva@polimi.it

Marta Spreafico, Davide Burba
MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, Italy, e-mail: marta.spreafico@polimi.it, e-mail: davide.burba@mail.polimi.it

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_19

representing their compensators (i.e., their predictable parts) as functional data. This approach originates from proposals in [2, 5]. In doing so, we end up with functional data representing the dynamic behavior of some covariates of interest, to be included into the model for the time-to-event process of interest (e.g., long-term survival). Suitable dimensionality reduction through Functional Principal Component Analysis (FPCA) [9] may then be carried out, and the resulting scores may be included in a predictive model.

We applied our methodology to analyse *HFData* administrative database provided by *Regione Lombardia - Healthcare Division* related to patients hospitalized for Heart Failure (HF) between 2000–2012 [7]. *HFData* provides, besides storing patients variables like age, gender and survival time, also observations related to drugs purchases and hospitalisations. Our analysis aims at providing a joint description of the HF related long-term survival and of the processes that may affect it (e.g., drugs purchase as a proxy of drugs intake, and hospitalisations).

This contribution is structured as follows: Sect. 19.2 presents the whole methodology, with a detailed description of the marked point process formulation for recurrent events. Sect. 19.3 reports the application to *HFData*. Finally, Sect. 19.4 contains some concluding remarks. All the analyses are carried out using the software R [10].

19.2 Methodology

In this section we focus on the main novelty introduced by the present work, i.e., the idea of representing the compensators of suitable marked point processes as functional covariates possibly affecting the outcome process of interest.

Let T_{start} be the time instant a HF patient is discharged by her/his first hospitalization and enrolled into the current study, and $T_0 = T_{start} + 365$ the starting time of the follow up. Moreover, let T_{end} be the minimum between the death or the administrative censoring (31-12-2012) for the same patient. The time-to-event process of interest, i.e., the long term survival of the patient, is measured on the time interval $T_{end} - T_0$. On the other hand, the compensators of the stochastic processes of interest is reconstructed on the time interval $T_0 - T_{start}$, called *observation period* in what follows.

Let's consider K recurrent events for a set of n individuals as stochastic processes. In particular, let's use *marked point process for recurrent events* [11], where a possibly multivariate *jump mark* $\mathbf{m}_i^{(k)}$ is associated to each *jump time* $t_i^{(k)}$. The observations (possibly censored) of multiple events for each individual may be seen as the realisation of an n-component multivariate counting process $\left(N_1^{(k)}, ..., N_n^{(k)}\right)$ where $N_i^{(k)}$ is the stochastic process which counts the observed events of the k-th process in the life of the i-th individual. According to the Doob-Meyer (D-M) decomposition [8], each counting process $N_i^{(k)}(t)$ is the sum of a martingale $M_i^{(k)}$, which represents the residual of the process, and a unique predictable increasing process $\Lambda_i^{(k)}(t) = \int_0^t \lambda_i^{(k)}(s)ds$. This predictable process, namely compensator, may be thought as a functionl datum, and will be the core of our modeling effort.
A counting process where jumps may have different size can be modelled as a point

process, assuming that a given distribution regulates the size of the jumps. A marked point process is then the couple of processes describing the behaviour of jumps and marks, and it is usually modelled through the *conditional intensity function* $\lambda^{(k)}(t, \mathbf{m}^{(k)} | \mathcal{F}_t^{(k)})$, i.e., the expected rate of events k at time t with mark $\mathbf{m}^{(k)}$:

$$\lambda^{(k)}\left(t, \mathbf{m}^{(k)} | \mathcal{F}_t^{(k)}\right) = \lambda_g^{(k)}\left(t | \mathcal{F}_t^{(k)}\right) f^{(k)}\left(\mathbf{m}^{(k)} | \mathcal{F}_t^{(k)}\right) \tag{19.1}$$

where k is the process of interest, $\mathcal{F}_t^{(k)}$ is the filtration of the process and it is interpreted as the history of realisations of the process, $\lambda_g^{(k)}$ is the *ground intensity*, i.e. the intensity process of the counting process, and $f^{(k)}$ is the multivariate density of the mark $\mathbf{m}^{(k)}$. Using this formulation, conditional independence of jump times and marks is assumed.

To handle recurrent events and allow predictors to change in time, we used the counting process formulation of the Cox model for recurrent events, as done in [1]. In particular, for each event k, the conditional intensity function $\lambda_i^{(k)}(t)$ of patient i in Eq. (19.1) takes the form:

$$\begin{aligned} \lambda_i^{(k)}(t) &= Y_i^{(k)}(t) \lambda_0^{(k)}(t) \exp\left\{\boldsymbol{\beta}^{(k)^T} \mathbf{x}_i^{(k)}(t)\right\} \exp\left\{\boldsymbol{\gamma}^{(k)^T} \mathbf{z}_i^{(k)}(t)\right\} \\ &= Y_i^{(k)}(t) \lambda_0^{(k)}(t) \exp\left\{\boldsymbol{\beta}^{(k)^T} \mathbf{x}_i^{(k)}(t) + \boldsymbol{\gamma}^{(k)^T} \mathbf{z}_i^{(k)}(t)\right\} \end{aligned} \tag{19.2}$$

where $\mathbf{x}_i^{(k)}(t)$ is the possibly time-dependent vector of covariates of the i-th individual, $\mathbf{z}_i^{(k)}(t)$ is the time-dependent vector of covariates related to the marks $\mathbf{m}_i^{(k)}$ of the i-th individual, $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$ are fixed vectors of coefficients, $\lambda_0^{(k)}$ is the underlying hazard function shared across patients, and $Y_i^{(k)}$ is a predictable process taking values in $\{0, 1\}$. Whenever $Y_i^{(k)} = 1$, the i-th individual is under observations (i.e. $Y_i^{(k)}$ takes the role of the censoring variable). The estimation of the parameters $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$ was based on a partial likelihood function [4], and maximised by applying the Newton-Raphson iterative procedure [6]. For all $k \in K$ the baseline cumulative hazard $\Lambda_0^{(k)}(t) = \int_0^t \lambda_0^{(k)}(s)ds$ can be estimated using the *Breslow estimator* $\hat{\Lambda}_0^{(k)}$ [3], which returned step-function. However, since true underlying functions $\Lambda_0^{(k)}$ are absolutely continuous, we smoothed the estimates using the approach adopted in [2], obtainining regularised version of $\Lambda_0^{(k)}$, namely $\tilde{\Lambda}_0^{(k)}$.

Then, we considered $t_{i,0}^{(k)} < t_{i,1}^{(k)} < \ldots < t_{i,N_i^{(k)}(\tau)}^{(k)}$ the realised jump times of process $N_i^{(k)}$, with τ equal to the censoring time (possibly equal for all individuals or not) and $t_{i,0}^{(k)} = 0$ for any k, i. We could express the realisations of each compensator $\Lambda_i^{(k)}$ for the k-th process of patient i as a function of $\Lambda_0^{(k)}$, $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$:

$$
\begin{aligned}
\Lambda_i^{(k)}(t) &= \int_0^t \lambda_i^{(k)}(s)ds = \int_0^t \lambda_0^{(k)}(s) e^{\boldsymbol{\beta}^{(k)^T} \mathbf{x}_i^{(k)}(s) + \boldsymbol{\gamma}^{(k)^T} \mathbf{z}_i^{(k)}(s)} ds \\
&= \sum_{j=1}^{N_i^{(k)}(t)} \int_{t_{i,j-1}^{(k)}}^{min\left(t_{i,j}^{(k)},t\right)} \lambda_0(s) e^{\boldsymbol{\beta}^{(k)^T} \mathbf{x}_i^{(k)}(t_{j-1}) + \boldsymbol{\gamma}^{(k)^T} \mathbf{z}_i^{(k)}(t_{j-1})} ds \\
&= \sum_{j=1}^{N_i^{(k)}(t)} e^{\boldsymbol{\beta}^{(k)^T} \mathbf{x}_i^{(k)}(t_{j-1}) + \boldsymbol{\gamma}^{(k)^T} \mathbf{z}_i^{(k)}(t_{j-1})} \left[\Lambda_0^{(k)} \left(min \left(t_{i,j}^{(k)}, t \right) \right) - \Lambda_0^{(k)} \left(t_{i,j-1}^{(k)} \right) \right]
\end{aligned}
\tag{19.3}
$$

An estimate of the compensator in Eq. (19.3) can be obtained as:

$$
\hat{\Lambda}_i^{(k)}(t) = \sum_{j=1}^{N_i^{(k)}(t)} e^{\hat{\boldsymbol{\beta}}^{(k)^T} \mathbf{x}_i^{(k)}(t_{j-1}) + \hat{\boldsymbol{\gamma}}^{(k)^T} \mathbf{z}_i^{(k)}(t_{j-1})} \left[\tilde{\Lambda}_0^{(k)} \left(min \left(t_{i,j}^{(k)}, t \right) \right) - \tilde{\Lambda}_0^{(k)} \left(t_{i,j-1}^{(k)} \right) \right]
\tag{19.4}
$$

where $\hat{\boldsymbol{\beta}}^{(k)}$ and $\hat{\boldsymbol{\gamma}}^{(k)}$ are the estimated vectors of coefficients and $\tilde{\Lambda}_0^{(k)}$ is the smoothed estimate of the cumulative baseline hazard. To check the fitting of $\hat{\Lambda}_i^{(k)}$, we graphically verified if the estimates $\hat{M}_i^{(k)}$ of martingale residuals $M_i^{(k)}$ involved in the D-M decomposition may be effectively considered as realisations of martingales observing if their means $\bar{M}^{(k)}(t) = \frac{1}{N}\sum_{i=1}^{N} \hat{M}_i{}^{(k)}(t)$ were approximately 0.

Applying this procedure for $k = 1, \ldots, K$, we end up with a $K-$variate functional data for each patient, characterizing her/his recurrent events dynamic in the observation period $T_0 - T_{start}$. We can now use such data in a methodological pipeline, as described in the next section.

19.3 Application and Results

We now present the analysis of *HFData*, obtained through a 4-steps procedure.

Step 1. Data preprocessing & Clinical history
We applied our methodology to a representative sample of *HFData* related to 4,872 patients with their first HF discharge between January 2006 to December 2012 [7]. The study-period started from the first discharge for HF (index date) and was divided into the *observation period* (365 days from the index discharge date) for the compensators reconstruction and the *follow-up period* for survival analysis. Only patients alive at the end of the observation period were followed up to observe survival outcomes. A final cohort of 4,541 (93.2%) patients was selected. Administrative censoring date was December 31^{st}, 2012.
We identified four types of stochastic processes of interest: rehospitalisations due to HF (hosp) and purchases of Angiotensin Converting Enzyme (ACE) inhibitors, Beta-Blocking (BB) agents and Anti Aldosterone (AA) agents, identified by their Anatomical Therapeutic Chemical (ATC) codes [12]. Hence, the set of recurrent events of interest was $K = \{k : ACE, BB, AA, HF\ hosp\}$. Finally, we selected only

events happened to patients within the one year observation period. Such events are named *"clinical history"* of the patients. Then, for each patient i, each event process was seen as a marked counting process $N_i^{(k)}$, with *jump times* $t_i^{(k)}$ equal to event times (i.e. date of admission in hospital or date of drug purchase) and *jump marks* $\mathbf{m}^{(k)}$ equal to the length of stay in hospital or the duration of drug coverage.

Step 2. Modeling compensators

2.1 *Features selection and coefficients estimation.* For each process k, we used as covariates $\mathbf{z}_i^{(k)}(t)$ of patient i: $Nm^{(k)}(t)$, i.e., the number of events related to the $k-$th process occurred in the past; $y^{(k)}(t)$, i.e., the sum of the corresponding marks. Also the logarithmic transformations (shifted away from 0) of the same variables, i.e., $log(Nm^{(k)}(t)+1)$ and $log(y^{(k)}(t)+1)$, and respective interactions, were considered. Adjustments for age and $gender$ at baseline were also performed. The vector of all the covariates considered for the model is indicated by $\mathbf{x}_i^{(k)}$.

Among all the models tested through a cross-bvalidation procedure, features related to $Nm^{(k)}(t)$, $y^{(k)}(t)$ and their interaction were selected and the signs of their fitted coefficients were consistent throughout the four processes, allowing us to give similar interpretations. In particular, $Nm^{(k)}(t)$ and $y^{(k)}(t)$ (or the corresponding logarithmic version) could be interpreted as indicators of a “self-exciting” behaviour. In other words, many drug purchases (or being hospitalized often in the past) and the purchase of big quantities of drug (or having spent long periods of time at the hospital) both increase the risk of a new purchase (or of a new hospitalisation) [HR>1]. Moreover, the interaction term being significant suggests that the increase in risk is softened in case of several drug purchases (or many hospitalizations) and/or a great quantities of drug purchased (or a long time spent at the hospital in the past) [HR<1]. Finally, we fitted four Cox models, one for each process k, using the selected features on the entire dataset to estimate coefficients $\hat{\boldsymbol{\beta}}^{(k)}$ and $\hat{\boldsymbol{\gamma}}^{(k)}$.

2.2 *Fit and smooth cumulative baseline hazard.* Once we estimated the coefficients $\hat{\boldsymbol{\beta}}^{(k)}$ and $\hat{\boldsymbol{\gamma}}^{(k)}$ of each of the k-th Cox model for recurrent events, we compute the estimated cumulative baseline hazards $\hat{\Lambda}_0^{(k)}$ with the Breslow estimator. Since this procedure provide a step function ($\hat{\Lambda}_0^{(k)}$), we smooth them using 20 knots spline basis, constraining $\tilde{\Lambda}_0^{(k)}(-0.5) = 0$. In doing so, for all $k \in K$ we obtain monotonically increasing estimates $\tilde{\Lambda}_0^{(k)}$ of the cumulative baseline hazards.

2.3 *Reconstruct compensators.* At this point, we can reconstruct the trajectories of the compensators $\hat{\Lambda}_i^{(k)}$ of the four considered stochastic processes for all the patients, exploiting Eq. (19.4). The trajectories of compensators $\hat{\boldsymbol{\Lambda}}_i^{(k)}$ constitute our time-varying covariates. Fig. 19.1 shows the compensators of the stochastic processes describing ACE purchase (top-left panel), BB purchase (top-right panel), AA purchase (bottom-left panel) and HF hospitalisation (bottom-right panel) of a sample of 500 HF patients belonging to *HFData*. Observe that the large variability of the compensators across different patients reflects the variability of the realizations

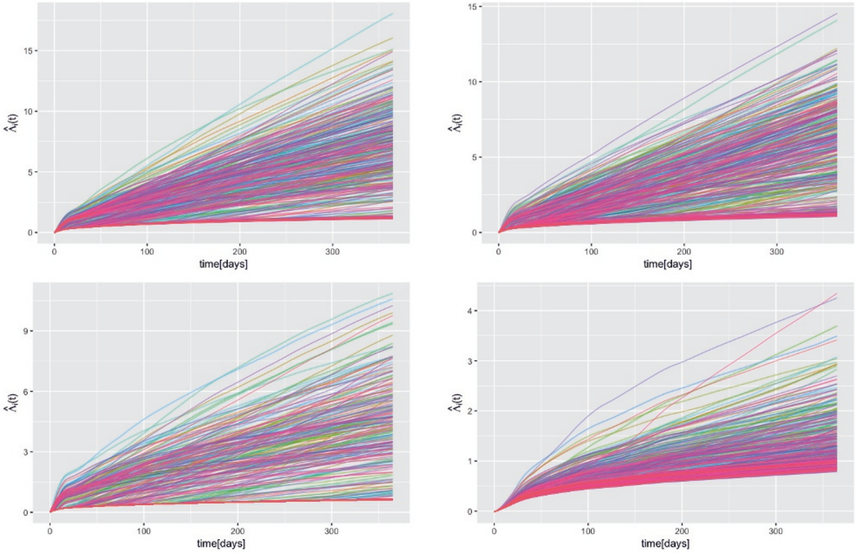


Fig. 19.1 Compensators of the marked counting processes of purchases of ACE inhibitors (top-left panel), BB agents (top-right panel), AA (bottom-left panel) and of HF hospitalisations (bottom-right panel) fitted using Eq. (19.4).

of their recurrent events. Finally, we checked that means of martingale residuals $\bar{M}^{(k)}(t)$ were approximately equal to 0 in order to check for adequate fitting of the procedure.

Step 3. Summarize compensators through FPCA

We then applied FPCA techniques to compensators estimated at the previous step. Since compensators are postive and increasing function, i.e., constrained functions, we apply FPCA only for summarizing information emerging from the time-varying compensators. Transformation approaches may be applied, before FPCA, but in this case do not drive to significant improvements in the final results. The first PCs distinguish patients with different risks: a patient with high scores on the first PC is likely to experience more events than a patient with a low score. The second PCs distinguish patients with different time distribution of the events: a patient with a high score on the second PC is likely to experience more events in the first months of the observation period and less events in the last months of the observation period than a patient with a low score.

Step 4. Predictive survival Cox's model

Appling 10-fold cross validation to select the best set of covariates among age, $gender$, first and second PCs of each group of compensators $\hat{\mathbf{\Lambda}}^{(k)}$, we selected the model with the highest Concordance Index, i.e.:

$$\lambda(t|\mathbf{x}) = \lambda_0(t)exp\{\beta_1 age + \beta_2 BB\_PC1 + \beta_3 hosp\_PC1 + \beta_4 hosp\_PC2\} \quad (19.5)$$

Then, we fitted the Cox's regression model with that choice of covariates on a training set composed by 70% of the data (3,179 patients).

All the covariates resulted statistically significant at confidence level 5%. Elder patients coherently have a higher risk of dying (HR = 1.065, 95% $CI = [1.0558; 1.0744]$). The HR relative to BB_PC1 was 0.997 (95% $CI = [0.9948; 0.9992]$), indicating that proper BB agents intake is correlated to longer life expectancy. The HR relative to $hosp\_PC1$ was 1.020 (95% $CI = [1.0080; 1.0323]$), standing as a proxy o patients' critical conditions (patients experiencing many hospitalizations in the past present a higher risk of dying). Interestingly, the HR relative to the $hosp\_PC2$ was 0.756 (95% $CI = [0.7039; 0.8119]$), meaning that patients with many hospitalizations at the beginning of the observation period and few hospitalizations in the end had higher survival probability, since they probably corresponded to the ones who had already experienced a critical phase of the disease and survived from it. Finally, we used the fitted model to predict survival time of patients belonging to the test set, i.e. the remaining 30% (1,362 patients), obtaining a Concordance Index equal to 67.56%.

19.4 Conclusions

In this work, we proposed an effective methodology to extract and summarize information from trajectories of compensators of suitable marked point processes for recurrent events intended as functional data. The development of this procedure is due to the need of effectively describing and resuming information from dynamic processes affecting an outcome process of interest, and to plug it into a Cox regression model. This methodology extends the one proposed in [2], allowing the counting processes to depend on their marks. Moreover, the introduction of this novel way to account for time-varying variables by means of compensators of marked stochastic processes, allowed for modeling self-exciting behaviors, for which the occurrence of events in the past increases the probability of a new event. This approach has three main advantages with respect to standard Cox regression models: first of all, it enables to properly exploitation of the role of repeated events within a time-to-event framework, usually accounted for using simple counts of events as fixed covariate. Second, it links FDA and survival analysis, enlarging the range of possible application of survival regression models. Finally, the application of the proposed predictive models to clinical data may allow to differentiate scheduling of controls according to predicted risk, modulating therapies according to patients behavior and enabling a real-time update of prognosis. This is not possible with standard Cox regression models, nor with other extensions of it.

References

[1] Andersen, P.K., Gill, R.D.: Cox's Regression Model for Counting Processes: A Large Sample Study. Ann. Stat. **10**(4), 1100–1120 (1982)

[2] Baraldo, S., Ieva, F., Paganoni, A.M. et al.: Outcome Prediction for Heart Failure Telemonitoring Via Generalized Linear Models with Functional Covariates. Scand. J. Stat. **40**(3), 403–416 (2013)

[3] Breslow, N.E.: Analysis of survival data under the proportional hazards model. Int. Stat. Rev., 45–57 (1975)

[4] Cox, D.R.: Regression models and life tables. J. Royal Stat. Soc. **34**(2), 187–220 (1972)

[5] Kennedy, B.: Repeated Hospitalizations and Self-rated Health among the Elderly: A Multivariate Failure Time Analysis. Am. J. Epidemiol. **153**, 232–241 (2001)

[6] Lee, E.T., Wang, J.: Statistical Methods for Survival Data Analysis. Wiley (2003)

[7] Mazzali, C., Paganoni, A.M., Ieva, F. et al.: Methodological issues on the use of administrative data in healthcare research: the case of heart failure hospitalizations in Lombardy region, 2000 to 2012. BMC Health Serv. Res. **16**(234), (2016)

[8] Meyer, P.A.: A decomposition theorem for supermartingales. Illinois J. Math. **6**, 193–205 (1962)

[9] Ramsay, J., Silverman, B.W.: Functional Data Analysis. Springer New York (2013)

[10] R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (2017) https://www.R-project.org/

[11] Ross, S.M.: Stochastic Processes. Wiley (1995)

[12] WHO Collaborating Centre for Drug Statistics Methodology: https://www.whocc.no

Chapter 20
On Robust Training of Regression Neural Networks

Jan Kalina and Petra Vidnerová

Abstract Estimation, prediction or smoothing of curves represents a fundamental task of functional data analysis. Nonlinear regression methods allow to search for the best-fit curves explaining the dependence of a response variable on available independent variables. Neural networks, commonly used for the task of nonlinear regression, are however highly vulnerable to the presence of outlying measurements in the data. New robust versions of common types of neural networks, namely multilayer perceptrons and radial basis function networks, are proposed here based on nonlinear regression quantiles or highly robust loss functions. Three datasets are analyzed to illustrate the performance of the novel robust approaches, which turn out to outperform standard neural networks or other competing regression tools over contaminated data.

20.1 Introduction

Analysis of curves (i.e. their modeling and/or smoothing) represents a fundamental task of functional data analysis [15, 20]. If the task is to model (explain and smoothen) the unknown shape of a response variable conditioning on the knowledge of available independent variables, one may use one of diverse tools of the nonlinear regression methodology, which is often denoted as function approximation. These numerous tools include various types of kernel smoothers, smoothing splines, or artificial neural networks. An important issue is however the performance of these tools of nonlinear regression under the presence of outlying measurements in the data.

Jan Kalina (✉)
The Czech Academy of Sciences, Institute of Computer Science, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic, e-mail: kalina@cs.cas.cz

Petra Vidnerová
The Czech Academy of Sciences, Institute of Computer Science, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic, e-mail: petra@cs.cas.cz

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_20

Regression neural networks allow to capture the multivariate structure of the regressors. The most commonly used methods for training regression neural networks based on the least squares criterion are biased under contaminated data [18] as well as vulnerable to adversarial examples. This non-robustness becomes even more severe for data with a very large number of regressors [4]. This is true for two important types of multilayer feedforward networks, namely multilayer perceptrons (denoted as MLPs) and radial basis function (RBF) networks. Therefore, researchers have recently become increasingly interested in proposing alternative robust (resistant) methods for their training [3, 10]. Only a few robust approaches for training for MLPs have been introduced. Some of them become computationally infeasible for a larger p [13] and thus were presented only on simple examples with a very small number p of regressors, Approaches replacing the common sum of squared residuals by a robust loss onsidered the loss functions corresponding to the median [2], least trimmed absolute value (LTA) estimator [17], or least trimmed squares (LTS) estimator [3, 18].

In this paper, we propose an original method for robust training neural networks based on nonlinear quantiles or robust loss functions. We also fill a gap of robust methods for other types of neural networks, mainly for RBF networks. Also, systematic comparisons of the performance of robust neural networks with other nonlinear regression tools seem to have been missing. Here, we compare the performance of various nonlinear regression tools including the newly proposed robust neural networks over three datasets.

20.2 Training Artificial Neural Networks

20.2.1 Model and Standard Approaches

MLPs and RBF neural networks represent well established classes of artificial neural networks suitable for the nonlinear regression model. Throughout the paper, we consider the regression task to model a continuous response $Y_1, \ldots, Y_n$ by means of p independent variables (regressors, features) available for n observations (measurements, instances), where the values for the i-th observation ($i = 1, \ldots, n$) are denoted as $X_{i1}, \ldots, X_{ip}$. We are interested in estimating the regression function

$$r(x) = \mathsf{E}(Y|X = x), \quad x \in \mathbb{R}^p, \tag{20.1}$$

based on the given data in the situation, when the shape of the function $r(x)$ is unknown and we only assume that it exists. We do not consider any assumptions on the shape of $r(x)$. Available estimates of (20.1) are often denoted as nonparametric, even if they depend on a finite number of parameters, because it is actually the whole function r which represents the unknown parameter. We will not formulate further assumptions; these would be however needed for deriving important properties of some of available estimators.

MLPs contain an input layer, one or more hidden layers with a fixed number of neurons, and an output layer. As we use the most standard form of MLPs, we will not present their detailed model, as it can be found in numerous monographs (see e.g. [7, 5]). Concerning RBF networks, the architecture of their most common type may be described as a hierarchical structure with an input layer containing p inputs, a single hidden layer with N RBF units (neurons), and a linear output layer. Most commonly [7, 10], the user chooses a radially symmetric function (kernel, basis function) denoted here as ψ.

Standard training of the most common types of neural networks, including MLPs and RBF networks, is based on minimizing the sum of squared residuals over all parameters, which are traditionally denoted as weights. The loss function in the form of sum of squared residuals causes the non-robustness of plain MLPs and RBF networks with respect to outlying measurements (outliers) [3, 17].

20.2.2 Inter-quantile Robust Neural Networks

We will now introduce regression τ-quantile estimators by means of MLPs and RBF networks, denoted as QMLP(τ) and QRBF(τ) for a fixed $\tau \in (0,1)$. Based on them, robust interquantile versions of neural networks will be defined in this section. Quantiles estimated by means of MLPs, also called quantile regression neural network, are available only for a single hidden layer (cf. [21]). They represent flexible tools for situations, when the regression curve remains unknown; thus, they may provide more complex information about the trend in the data compared to (20.1).

The novel QRBF(τ) estimator is formally defined by replacing the usual quadratic loss function by ρ_τ (see [11])

$$\rho_\tau(x) = x\,(\tau - \mathbb{1}[x < 0])\,, \quad x \in \mathbb{R}, \tag{20.2}$$

with indicator function denoted by $\mathbb{1}$. Denoting fitted values of the response by $\hat{Y}_i$, the computation requires to find

$$\min \sum_{i=1}^{n} \rho_\tau\left(Y_i - \hat{Y}_i\right) \tag{20.3}$$

over the space of networks parameters.

We define the new interquantile robust RBF networks denoted as IQ-RBF networks, depending on two given parameters τ_1 and τ_2, by means of Algorithm 1. Interquantile MLPs (IQ-MLP) are defined in an analogous way. The network is trained for observations between two different quantiles with parameters τ_1 and τ_2; the outlier detection and deletion (trimming) by means of the quantiles ensure the resistance against outliers. The approach is inspired by the trimmed least squares (TLS) estimator in linear regression, which has appealing robustness properties (see [8] for their extensive discussion).

Algorithm 1 IQ-RBF network.

```
Input: Response $Y_1, \ldots, Y_n$
Input: Regressors $X_i \in \mathbb{R}^p$ for $i = 1, \ldots, n$
Input: Constants $\tau_1$ and $\tau_2$, where $\tau_1 \in (0,1)$, $\tau_2 \in (0,1)$, and $\tau_1 < 1/2 < \tau_2$
Output: IQ-RBF trained with parameters $\tau_1$ and $\tau_2$
Compute QRBF($\tau_1$) network
For each $i = 1, \ldots, n$, put $\hat{Y}_i^{\mathsf{QRBF}(\tau_1)}$ equal to the fitted value of $Y_i$ by QRBF($\tau_1$) network
Compute QRBF($\tau_2$) network
For each $i = 1, \ldots, n$, put $\hat{Y}_i^{\mathsf{QRBF}(\tau_2)}$ equal to the fitted value of $Y_i$ by QRBF($\tau_2$) network
Fit a standard RBF network only for such measurements, for which
    $$Y_i \geq \hat{Y}_i^{\mathsf{QRBF}(\tau_1)} \quad \& \quad Y_i \leq \hat{Y}_i^{\mathsf{QRBF}(\tau_2)} \qquad (20.4)$$
```

20.2.3 Neural Networks with a Robust Loss Function

The MLP with the LTA-based loss function [17, 18] will be now denoted as LTA-MLP. It is defined for a fixed h ($n/2 \leq h < n$) by means of

$$\arg\min_{b \in \mathbb{R}^p} \sum_{i=1}^{h} |u(b)|_{(i)}. \tag{20.5}$$

In [17], LTA-MLP yielded slightly better results than LTS-MLP, which seems inspiring for the field of robust statistics, where the LTA estimator is practically unknown. The LTS-MLP, based on the least trimmed squares (LTS) estimator (see [8]), is defined by

$$\arg\min_{b \in \mathbb{R}^p} \sum_{i=1}^{h} u_{(i)}^2(b). \tag{20.6}$$

Robust RBF networks, denoted here as LTA-RBF or LTS-RBF networks, will be defined by means of (20.5) and (20.6) in the context of RBF networks. In other words, they are obtained by replacing the usual quadratic loss (i.e. sum of squared residuals) by the loss functions of the LTS or LTA estimators.

20.3 Numerical Examples

Three datasets will be now analyzed to illustrate the performance of the novel methods for nonlinear regression. All three have been found as suitable for regression modeling in the literature. In order to improve the convergence of the neural networks training, each dataset is considered after a usual standardization, i.e. each variable is centered to have the mean equal to 0 and scaled to have variance equal to 1.

The methods used in all the computations include the Nadaraya-Watson (N-W) kernel regression with Epanechnikov kernel and regularized networks with a Gaussian kernel [7]. MLPs are always used with a sigmoid activation function in the hidden layer (or layers) and a linear output layer. For RBF networks, we choose ψ to be the Gaussian density. Other parameters of MLPs and RBF networks were chosen to obtain the best possible results (in terms of MSE) in every example, while the robust versions have the same parameters as the plain (standard) versions. Methods based on the LTS and LTA use the trimming constant $h = \lfloor 3n/4 \rfloor$, where $\lfloor x \rfloor$ denotes the integer part of $x \in \mathbb{R}$. Interquantile methods (IQ-MLP and IQ-RBF) use $\tau_1 = 0.15$ and $\tau_2 = 0.85$. Table 1 presents for each of the regression methods a reference (or it reveals which methods are novel) together with a reference for the software code for the computation.

We use a standard back-propagation algorithm for the task of minimisation to obtain all neural networks computed here. Particularly, the RMSprop optimization technique was used. All robust versions of MLPs and RBF networks were implemented exploiting the TensorFlow library [1] of Python. We implemented regularized networks in R. The results are evaluated in a 10-fold cross validation for all three datasets.

We consider two versions of the prediction error, namely the standard mean squared error (MSE) computed over all observations, or its robust counterpart denoted as the trimmed mean squares error (TMSE). Using $\alpha = 3/4$ and $h = \lfloor \alpha n \rfloor$, these are defined as

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} u_i^2 \quad \text{and} \quad \text{TMSE}(\alpha) = \frac{1}{h}\sum_{i=1}^{h} u_{(i)}^2, \tag{20.7}$$

where the prediction errors $u_1, \ldots, u_n$ are considered after being arranged in ascending order (in squares) as $u_{(1)}^2 \leq u_{(2)}^2 \leq \cdots \leq u_{(n)}^2$.

20.3.1 Results

The first dataset is the Travel and Tourism Competitiveness Index (TTCI) dataset with $p = 12$ and $n = 141$, which was previously analyzed by (robust) linear regression methods in [9]. Numerical values of the prediction errors for various nonlinear tools are presented in Table 1 here. A standard MLP with 12 neurons in a single hidden layer is influenced by the presence of outliers, which is true also for a plain RBF network with $N = 12$ RBF units. While the novel methods do not improve MSE compared to standard ones (and are not expected to improve a non-robust version of prediction error), they improve values of TMSE.

The second dataset is the Boston Housing dataset [6] with the per capita crime rate in different Boston suburbs as the response variable. We selected all $p = 11$ continuous features (omitting features 4, 7, and 9 from the original dataset) to be the regressors. There are no missing values in the dataset, which allows us to work with all $n = 506$ measurements. We use MLPs containing 16 and 8 neurons in the two hidden layers and RBF networks with $N = 50$. Values of prediction error are compared in Table 1. While the novel methods do not improve MSE, they improve values of TMSE. This corresponds to our empirical evidence that the Boston Housing dataset contains about 10 % of severe outliers.

		Dataset		
		TTCI	Boston housing	Auto MPG
Method	Source/Code	MSE/TMSE	MSE/TMSE	MSE/TMSE
N-W	[19]/R	0.51/0.17	54.3/4.3	53.7/21.3
Reg. networks	[7]/Own	0.47/0.15	60.6/4.7	61.0/19.4
		Versions of MLP		
MLP	[7]/[1]	0.41/0.14	57.9/5.3	60.8/28.9
RMLP	[10]/Own	0.44/0.12	65.1/4.3	72.8/15.0
LTS-MLP	[17]/Own	0.43/0.12	67.2/4.5	69.4/14.3
LTA-MLP	[17]/Own	0.43/0.12	66.8/4.5	69.6/14.1
IQ-MLP	Novel/Own	0.44/0.12	67.7/4.2	70.1/13.8
		Versions of RBF network		
RBF	[7]/[1]	0.39/0.14	52.7/4.4	46.9/17.2
RRBF	[10]/Own	0.43/0.12	59.7/3.9	51.0/13.3
LTS-RBF	Novel/Own	0.45/0.12	60.3/4.1	52.7/12.9
LTA-RBF	Novel/Own	0.45/0.12	61.1/4.1	53.2/12.7
IQ-RBF	Novel/Own	0.44/0.11	60.8/3.7	52.3/12.2

Table 20.1 Results for the three datasets of Section 20.3 evaluated in a 10-fold cross validation. Prediction error measures (MSE and TMSE($3/4$)) evaluated for various (standard or robust) nonlinear regression methods.

The third dataset is the Auto MPG dataset [6] with the consumption of cars in miles per gallon (MPG) playing the role of the response, explained by $p = 4$ continuous regressors, namely displacement, horsepower, weight, and acceleration. Observations with some missing values (i.e. observations with index 33, 127, 331, 337, 355, and 375) are omitted, so we work with $n = 392$. For the analysis, we use MLPs with 16 and 8 neurons in the hidden layers and RBF networks with $N = 40$. The results are again presented in Table 1. The novel methods do not improve MSE but do improve TMSE compared to standard methods. This argument in accordance with the fact that the Auto MPG dataset is contaminated by severe outliers as well as leverage points.

20.4 Conclusions

Regression neural networks, commonly used for modeling the unknown shape of the relationship of a response variable on available regressors, are vulnerable to the presence of outliers in the data. The effect of outliers becomes more intricate with an increasing number of regressors in the model and even more if the number of regressors exceeding the number of observations, i.e. in situations with $n < p$ [14]. In this paper, several novel approaches for robust training of MLPs and RBF networks, i.e. two very common types of neural networks, are proposed for the task of nonlinear regression. The computation of all the novel methods is straightforward and can rely on adapting standard optimization algorithms. Properties of the novel robust methods depend on choosing a particular architecture and parameters, which is supported by numerical evidence.

The numerical examples of this paper illustrate the novel robust neural networks to be meaningful over three real datasets. The results reveal the non-robustness of standard (plain) tools and the robustness of the novel methods based on quantiles (IQ-MLP, IQ-RBF) or robust loss functions (LTS-RBF, LTA-RBF). If suitable architecture for the neural networks was found and considered, the robust neural networks yield a smaller robust prediction error compared to other methods over the contaminated data. In other words, the robust methods are able to outperform standard neural networks or other (non-robust) regression tools including also the popular Nadaraya-Watson kernel regression. The computation of the novel approach is far less demanding compared to robust neural networks of [3, 10], which are computationally very tedious already for 10 regressors.

Neural networks (in their plain form) have also been successfully applied to functional data in various research fields (see e.g. [12]), while versions of neural networks adapted for the context of functional data are also available [16]; the latter were obtained as an extension of traditional neural networks for regression to the situation with $p \to \infty$ [13, 14]. As the next natural step, we would like to extend the robust training proposed in this paper to convolutional neural networks, or to combine robustness with regularization; the performance of such novel approaches to training neural networks is intended to be investigated over high-dimensional or functional data.

Acknowledgements The research was supported by the projects GA19-05704S and GA18-23827S of the Czech Science Foundation. The authors are grateful to Jan Tichavský for technical help.

References

[1] Abadi, M. et al.: Tensorflow: A system for large-scale machine learning. (2015) http://tensorflow.org

[2] Aladag, C.H., Egrioglu, E., Yolcu, U.: Robust multilayer neural network based on median neuron model. Neural Computing and Applications **24**, 945–956 (2014)

[3] Beliakov, G., Kelarev, A., Yearwood, J.: Derivative-free optimization and neural networks for robust regression. Optimization **61**, 1467–1490 (2012)

[4] Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., Usunier, N.: Parseval networks: Improving robustness to adversarial examples. Proceedings of the 34th International Conference on Machine Learning ICML 2017, 854–863 (2017)

[5] Du, K.L., Swamy, M.N.S.: Neural networks and statistical learning. Springer, London (2014)

[6] Frank, A., Asuncion, A.: UCI Machine Learning Repository. University of California, Irvine (2010) http://archive.ics.uci.edu/ml

[7] Haykin, S.O.: Neural networks and learning machines: A comprehensive foundation. 2nd edn. Prentice Hall, Upper Saddle River (2009)

[8] Jurečková, J., Picek, J., Schindler, M.: Robust statistical methods with R. 2nd edn. CRC Press, Boca Raton (2019)

[9] Kalina, J., Vašaničová, P., Litavcová, E.: Regression quantiles under heteroscedasticity and multicollinearity: Analysis of travel and tourism competitiveness. Ekonomický časopis **67**, 69–85 (2019)

[10] Kalina, J., Vidnerová, P.: Robust training of radial basis function neural networks. Proceedings 18th International Conference ICAISC 2019, 113–124 (2019)

[11] Koenker, R.: Quantile regression: 40 years on. Annual Review of Economics **9**, 155–176 (2017)

[12] Masselot, P., Dabo-Niang, S., Chebana, F., Ouarda, T.B.M.J.: Streamflow forecasting using functional regression. Journal of Hydrology **538**, 754–766 (2016)

[13] Pang, G., Chen, L., Cao, L., Liu, H.: Learning representations of ultrahigh-dimensional data for random distance-based outlier detection. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD'18), 2041–2050 (2018)

[14] Popovic, D., Fouché, E., Böhm, K.: Unsupervised artificial neural networks for outlier detection in high-dimensional data. Lecture Notes in Computer Science **11695**, 3–19 (2019)

[15] Ramsay, J., Silverman, B.W.: Functional data analysis. 2nd edn. Springer, New York (2005)

[16] Rossi, F., Delannay, N., Conan-Guez, B., Verleysen, M.: Representation of functional data in neural networks. Neurocomputing **64**, 183–210 (2007)

[17] Rusiecki, A.: Robust learning algorithm based on LTA estimator. Neurocomputing **120**, 624–632 (2013)

[18] Rusiecki, A., Kordos, M., Kamiński, T., Greń, K.: Training neural networks on noisy data. Lecture Notes in Artificial Intelligence **8467**, 131–142 (2014)

[19] Shawe-Taylor, J., Cristianini, N.: Kernel methods for pattern analysis. Cambridge University Press, Cambridge (2004)

[20] Ullah, S., Finch, C.F.: Applications of functional data analysis: A systematic review. BMC Medical Research Methodology **13**, Article 43 (2013)

[21] Xu, Q., Deng, K., Jiang, C., Sun, F., Huang, X.: Composite quantile regression neural network with applications. Expert Systems with Applications **76**, 129–139 (2017)

Chapter 21
Simultaneous Inference for Function-valued Parameters: a Fast and Fair Approach

Dominik Liebl and Matthew Reimherr

Abstract This work presents a new approach for constructing simultaneouse confidence bands for function-valued parameters. The bands are fast to compute as they are based on nearly closed-form expressions and, therefore, do not require computationally expensive resampling based methods. The shape of the bands can be constructed according to a desired criteria specified by the user. A particularly interesting criteria is the proposed concept of "fair" or equitable bands which leads to simultaneous confidence bands that have an adaptive width reflecting the local multiple testing problem. The theoretical foundations of our simultanouse confidence bands are presented in [9]. In this short paper, we deviate from [9] and consider the practically important special case of the linear function-on-scalar regression model. Moreover, we present a novel application on testing for differences in yield curves of A and B-type rated countries.

21.1 Introduction

This work is concerned with constructing simultaneous confidence bands that allow to quantify estimation uncertainty for function-valued parameters. While the theoretical foundations of our bands are presented in [9], we focus in this paper on the applications of our bands to the practically important case of the function-on-scalar regression model.

Simultaneous inference for function-valued parameters is a highly relevant statistical problem and has received increasing attention in recent years. However, current solutions usually suffer from some major drawbacks: they are computationally ex-

Dominik Liebl (✉)
Institute of Finance and Statistics and Hausdorff Center for Mathematics, University of Bonn, Adenauerallee 24-26, 53113 Bonn, Germany, e-mail: dliebl@uni-bonn.de

Matthew Reimherr
Department of Statistics, Penn State University, 411 Thomas Building University Park, PA 16802, U.S.A., e-mail: mreimherr@psu.edu

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_21

pensive, they result in bands that are very conservative, or they control only the point-wise type-I error rate.

By contrast to the majority of existing alternative approaches (see, for instance, [4, 5, 1, 6, 10, 3, 16]), our bands do not rely on re-sampling or simulation-based methods which makes our bands fast to compute. Building upon results from random field theory (see [2]) we propose a method that exploits the expected Euler characteristic inequality and derives a simultaneous confidence band using the Kac-Rice formula (see [7, 12]). We extend the existing Kac-Rice formula in order to allow for band-shapes that adapt to a desired criteria such a minimum width or equitable inference (fair). Related to our approach is the work of [14] who also propose a random field theory based confidence band; this approach, however, does not consider the case of adaptive band shapes.

The rest of the paper is structured as follows. In Section 21.2 we present our methodological and theoretical contributions, Section 21.3 contains our real-data application and Section 21.4 concludes.

21.2 Simultaneous Confidence Bands

Let us consider the linear function-on-scalar (see, e.g., Ch. 13 in [11]) regression model

$$y_i(t) = x_i\boldsymbol{\beta}(t) + \varepsilon_i(t), \quad t \in [0,1], \quad i = 1,\dots,n, \tag{21.1}$$

with continuous dependent variable $y_i \in C^1([0,1])$, deterministic or random predictors $x_i = (1, x_{i2},\dots,x_{iK}) \in \mathbb{R}^K$, $1 \le K < \infty$, unknown continuous parameter functions $\boldsymbol{\beta}(t) = (\beta_1(t),\dots,\beta_K(t))^\top$, with $\beta_j \in C^1([0,1])$ for $j = 1,\dots,K$, and continuous random error function $\epsilon_i \in C^1([0,1])$. We consider the case of an iid sample $(Y_i, x_i, \varepsilon_i)$ distributed as (Y, x, ε) for all $i = 1,\dots,n$, where ε is independent from the predictors x. The error function ε is assumed to be a Gaussian processes with zero mean function $\mathbb{E}(\varepsilon(t)) = 0$ for all $t \in [0,1]$ and a finite, non-zero, covariance function $\sigma(s,t) = \mathbb{E}(\varepsilon(s)\varepsilon(t))$ with $t, s \in [0,1]$. All statements above involving stochastic quantities such as $y_i \in C^1([0,1])$ are meant to hold almost surely.

For reasons of traceability, we focus in this short paper on the special case of a Gaussian error process ε with known covariance function σ. The practically more relevant case of a unknown covariance function is considered in [9]. Based on the central limit theorems in [6], one can also relax the normality assumption (see also our theoretical paper [9]).

Under the above setup, the unknown functional slope coefficients, $\boldsymbol{\beta}(t) = (\beta_1(t),\dots,\beta_K(t))^\top$, in model (21.1) can be consistently estimated using the ordinary least squares estimator

$$\begin{aligned}\widehat{\boldsymbol{\beta}}(t) &= (\widehat{\beta}_1(t),\dots,\widehat{\beta}_K(t))^\top = \left(X^\top X\right)^{-1} X^\top Y(t), \quad \text{where}\\ \widehat{\beta}_j(t) &\sim \mathcal{N}\big(\beta_j(t), \sigma(t,t)\big[(X^\top X)^{-1}\big]_{jj}\big), \quad j = 1,\dots,K,\end{aligned}$$

with X denoting the $n \times K$ predictor matrix $[X]_{ij} = x_{ij}$ and $Y(t) = (y_1(t), \ldots, y_n(t))^\top$. If X contains stochastic components, the above distributional statement is understood as conditionally on X; this convention is used also in the remaining parts of this paper.

In order to construct a simultaneous confidence band for β_j, $j = 1, \ldots, K$, that allows for an adaptive band-shape, we allow for a generally non-constant, continuous critical value function $c_\alpha(t) > 0$ such that

$$P\Big(\beta_j(t) \in \widehat{\beta}_j(t) \pm c_{\alpha,j}(t)\sqrt{\operatorname{Var}\big(\widehat{\beta}_j(t)\big)}\ \forall\, t \in [0,1]\Big) \geq 1-\alpha, \quad j = 1, \ldots, K,$$

where α denotes the significance level (for instance, $\alpha = 0.05$). Using the symmetry of the distribution of $\widehat{\beta}_j(t)$, one can define the critical value function, $c_\alpha(t)$, by the following equation

$$\begin{aligned} &P\big(\exists t \in [0,1] : Z_j(t) \geq c_{\alpha,j}(t)\big) \leq \alpha/2, \quad \text{where} \qquad (21.2)\\ &Z_j(t) = \frac{\big(\widehat{\beta}_j(t) - \beta_j(t)\big)}{\sqrt{\sigma(t,t)\big[(X^\top X)^{-1}\big]_{jj}}} \sim \mathcal{N}(0,1), \quad j = 1, \ldots, K. \end{aligned}$$

That is, Z_j is a mean-zero Gaussian process $Z_j \sim \mathcal{GP}(0, \sigma[(X^\top X)^{-1}]_{jj})$, $j = 1, \ldots, K$.

Unfortunately, finding the exact, i.e., non-conservative critical value function, $c_\alpha(t)$, for which the left-hand-side of Equation (21.2) equals $\alpha/2$, is a very tricky problem since we do not want to impose restrictive assumptions on the structure of the covariance function σ. Therefore, we propose to derive a slightly conservative critical value function by making use of the so-called expected Euler characteristic inequality (see, for instance, [2]). This inequality involves the following random counting variable, $N_{u,Z_j}([0,1])$, which counts the number of up-crossings of Z_j about c_α on the interval $[0,1]$,

$$N_{c,Z_j}([0,1]) := \#\{0 \leq t \leq 1 : Z_j(t) = c(t), Z_j'(t) > c'(t)\}, \quad j = 1, \ldots, K.$$

If $N_{c,Z_j}([0,1]) = 0$, then the only way that $Z_j(t)$ could have exceeded $c(t)$ was if Z_j started above of c at $t = 0$, since both functions, Z_j and c, are continuous. This leads to the expected Euler characteristic inequality which follows directly from applying Boole's inequality and then Markov's inequality:

$$\begin{aligned} P\big(\exists t \in [0,1] : Z_j(t) \geq c(t)\big) &= P\big(\{Z_j(0) \geq c(0)\} \text{ or } \{N_{c,Z_j}([0,1]) \geq 1\}\big)\\ &\leq P\big(Z_j(0) \geq c(0)\big) + P\big(N_{c,Z_j}([0,1]) \geq 1\big)\\ &\leq P\big(Z_j(0) \geq c(0)\big) + \mathbb{E}\big[N_{c,Z_j}([0,1])\big] = \mathbb{E}\big[\varphi_c(Z_j)\big], \end{aligned} \tag{21.3}$$

for $j = 1, \ldots, K$, where $\varphi_c(Z_j) := \mathbb{1}_{Z_j(0) \geq c(0)} + N_{c,Z_j}([0,1])$ denotes the Euler (or Euler-Poincaré) characteristic of the excursion set $\{t \in [0,1] : Z(t) \geq c(t)\}$. For large values c, the inequality in (21.3) gives a very tight, almost exact approximation (see the results in [13]). Intuitively, the Euler characteristic, $\varphi_c(Z_j)$, counts exceedance events by, first, checking if $Z_j(0) > c(0)$ and then by checking for additional exceedance event by moving t through the domain $[0, 1]$.

The next step is to derive an expression for the mean of the Euler characteristic, $\mathbb{E}[\varphi_c(Z_j)]$, since this allows one to derive powerful simultaneous confidence bands by finding the critical value function $c_\alpha(t)$ which solves

$$\mathbb{E}[\varphi_{c_\alpha}(Z_j)] = \mathbb{E}\left[\mathbb{1}_{Z_j(0)\geq c(0)}\right] + \mathbb{E}\left[N_{c,Z_j}([0,1])\right] = \alpha/2.$$

The crucial, non-trivial part is to derive an expression for the mean of the counting variable $N_{c,Z}([0, 1])$. Explicit formulas for $\mathbb{E}[N_{c,Z}([0, 1])]$ are grouped together under the famous “Kac-Rice formulas” acknowledging the works of [7] and [12]. In [9] we derive a generalized Kac-Rice formula that allows us to consider adaptive critical value functions $c_\alpha(t)$. In the following we present the corresponding result in [9] for the special case of the function-on-scalar regression model (21.1) with Gaussian error processes. The required theoretical assumptions are as following.

Assumption (A1)

a) $Z_j = \{Z_j(t) : t \in [0, 1]\}$ is a centered Gaussian process, $Z_j \sim \mathcal{GP}(0, \sigma_{Z_j})$, where

$$\sigma_{Z_j}(s,t) = \frac{\sigma(s,t)[(X^\top X)^{-1}]_{jj}}{\sqrt{\sigma(s,s)[(X^\top X)^{-1}]_{jj}\;\sigma(t,t)[(X^\top X)^{-1}]_{jj}}}, \quad j = 1, \dots, K.$$

b) $Z_j \in C^1[0, 1]$ almost surely.
c) The covariance function σ_{Z_j} of Z_j is such that $\sigma_{Z_j}(s, t) = 1$ if and only if $s = t$.
d) Define the roughness function $\tau_j(t) = \sqrt{\mathrm{Var}(Z_j'(t))}$ and let $\tau_j(t) > 0$. □

Points a) and b) are fulfilled under our above introduced setup for the function-on-scalar regression model (21.1). Point c) is essentially a structural regularity assumption on the covariance function $\sigma(s,t)$ of the error processes ε in model (21.1) and excludes the case of periodic processes ε with $\varepsilon(t_1) = \varepsilon(t_2)$, almost surely, for some $t_1 \neq t_2$. Note that the assumption in point c) is by far less restrictive than, for instance, the assumption of a stationary covariance function as used by [15]. The roughness function $\tau_j(t)$ in point d) allows to measure the dependence structure of Z_j which is important to quantify the multiple testing problem that needs to be considered by the simultaneous confidence band.

Under Assumption (A1) one can derive the following result.

Theorem 1 *Let Assumption A1 hold and assume that the critical value function* $c \in C^1[0, 1]$, *then we have that*

$$\mathbb{E}[\varphi_c(Z_j)] = \Phi\left(-c(0)\right) + \int_0^1 \frac{\tau(t)}{2\pi} \exp\left\{-\frac{1}{2}\left[c(t)^2 + \frac{c'(t)^2}{\tau(t)^2}\right]\right\} dt$$
$$+ \int_0^0 \frac{c'(t)}{\sqrt{2\pi}} \exp\left\{-\frac{c(t)^2}{2}\right\} \Phi\left(\frac{c'(t)}{\tau(t)}\right) dt, \qquad j = 1, \ldots, K. \tag{21.4}$$

See [9] for a proof of a generalized version of Theorem 21.4. Theorem 21.4 allows for non-constant critical value functions and, therefore, generalizes the classic Gaussian Kac-Rice formula. In fact, for a constant critical value, $c(t) \equiv c$, Equation (21.4) simplifies to the case of the classic Gaussian version of the Kac-Rice formula, namely: $\mathbb{E}[\varphi_c(Z_j)] = \Phi(-c) + \int_0^1 \tau(t)/(2\pi)\exp\{-c^2/2\}dt$.

Equation (21.4) can now be used to determine a critical value function $c_\alpha(t)$ by solving $\mathbb{E}[\varphi_{c_\alpha}(Z_j)] = \alpha/2$. However, the this equation is generally solved by many different critical value functions c_α. That is, there are many different band-shapes that lead to a valid simultaneous $(1-\alpha) \times 100\%$ confidence band. This may seem to be an unfavorable situation, but it is more of a blessing than a curse since it allows us to select band-shapes according to user specified criteria such as minimum width or fair band width that reflects the local multiple testing problem. A detailed discussion of the computational methods for solving $\mathbb{E}[\varphi_{c_\alpha}(Z_j)] = \alpha/2$ by taking into account user specified criteria is beyond the scope of this short paper, but the interested reader is referred to [9].

21.3 Application

In this section we demonstrate the applicability of the above introduced simultaneous confidence for the function-on-scalar regression model. The yield-curve data shown in Figure 21.1 is available at www.worldgovernmentbonds.com.

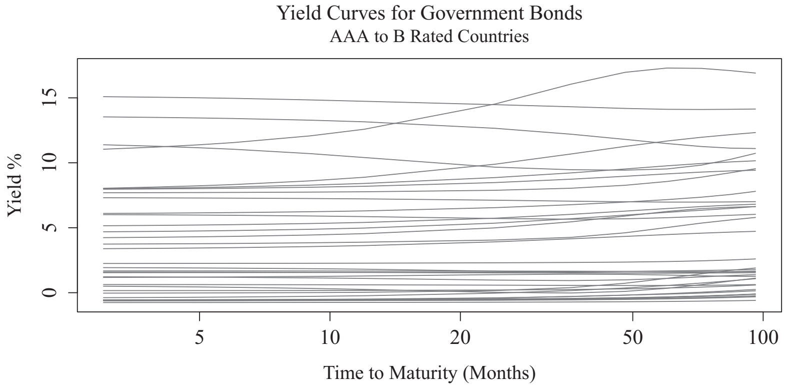


Fig. 21.1 Yield curves data for $n = 36$ government bonds for Jan 13, 2020.

In order to test for a difference in the means of A-type rated countries and B-type rated countries, we use the following linear function-on-scalar regression model

$$y_i(t) = \beta_1(t) + \beta_2(t)x_i + \varepsilon_i(t),$$

where y_i denotes the yield-curve of country i, with $i = 1, \dots, n = 36$, x_i is a dummy variable with $x_i = 0$ if country i is a A-type (AAA/AA/A) rated country and $x_i = 1$ if country i is a B-type (BBB/BB/B) rated country. The null-hypothesis of no difference between a A-type rating and a B-type rating is then formalized as

$$\mathrm{H}_0\colon \beta_1(t) = 0 \text{ simultaneously for all } t \in [0, 1].$$

The estimation result $\widehat{\beta}_1$ and the simultaneous 95% confidence interval for β_1 is shown in Figure 21.2 and we can conclude that $\widehat{\beta}_1$ is significantly different from zero everywhere on the domain.

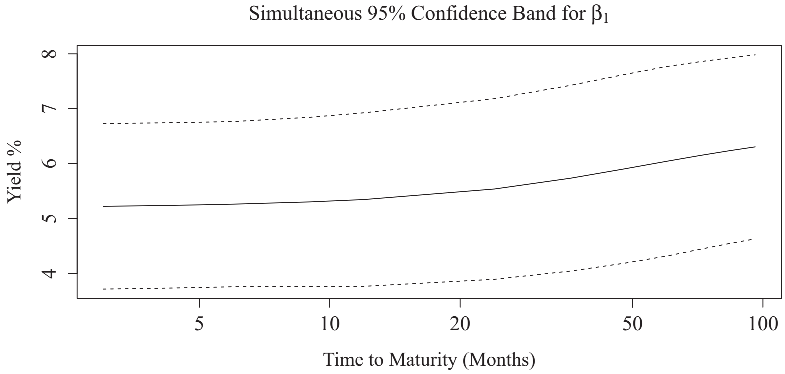


Fig. 21.2 Simulations 95% confidence band for the function-valued parameter β_1 (dashed lines) and the estimate $\widehat{\beta}_1$ (solid line).

21.4 Conclusion

This short paper derives simultaneous confidence bands for the function-valued coefficients in function-on-scalar regression models using the theoretical results in [9]. Our derivations focus on the simple case of Gaussian error processes ε, but may also be derived for elliptical processes following the more general results in [9]. We use a pointwise estimation approach for estimating the parameter functions $\boldsymbol{\beta}(t) = (\beta_1(t), \dots, \beta_K(t))^\top$ which is justified in our application since the yield-curves are very smooth. In the case of irregular sampled noisy response functions, one would typically prefer some regularized estimation approach such as, for instance, a penalized least squares estimation (see, for instance, [8], Ch. 5.2). Therefore, it would be useful to extend our work for regularized estimation procedures.

References

[1] Abramowicz, K., Häger C.K., Pini A., Schelin L., Sjöstedt de Luna S., Vantini S.: Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament. Scandinavian Journal of Statistics **45**(4), 1036–1061 (2018)

[2] Adler, R.J., Taylor J.E.: Random Fields and Geometry, 1st ed. Springer (2007)

[3] Cao, G., Yang, L., Todem, D.: Simultaneous inference for the mean function based on dense functional data. Journal of Nonparametric Statistics **24**(2), 359–377 (2012)

[4] Chang, C., Ogden, R.T. Bootstrapping sums of independent but not identically distributed continuous processes with applications to functional data. Journal of Multivariate Analysis **100**(6), 1291–1303 (2009)

[5] Degras, D.A.: Simultaneous confidence bands for nonparametric regression with functional data. Statistica Sinica **21**(4), 1735–1765 (2011)

[6] Dette, H., Kokot, K., Aue, A.: Functional data analysis in the Banach space of continuous functions. The Annals of Statistics, forthcoming (2020)

[7] Kac, M.: On the average number of real roots of a random algebraic equation. Bulletin of the American Mathematical Society **49**(4), 314–320 (1943)

[8] Kokoszka, P., Reimherr, M.: Introduction to Functional Data Analysis, 1st ed. Chapman & Hall/CRC (2017)

[9] Liebl, D., Reimherr, M.: Fast and Fair Simultaneous Confidence Bands for Functional Parameters. arXiv:1910.00131 (2020)

[10] Lopes, M.E., Lin, Z., Müller, H.G.: Bootstrapping max statistics in high dimensions: Near-parametric rates under weak variance decay and application to functional data analysis. The Annals of Statistics, forthcoming (2020)

[11] Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, 2nd ed. Springer (2005)

[12] Rice, S.O.: Mathematical analysis of random noise. Bell System Technical Journal **24**, 46–156 (1945)

[13] Taylor, J., Takemura, A., Adler, R.J.: Validity of the expected euler characteristic heuristic. The Annals of Probability **33**(4), 1362–1396 (2005)

[14] Telschow, F.J.E., Schwartzman, A.: Simultaneous confidence bands for functional data using the gaussian kinematic formula. arXiv:1901.06386 (2020)

[15] Wang, J., Cao, G., Wang, L., Yang, L.: Simultaneous Confidence Band for Stationary Covariance Function of Dense Functional Data. Journal of Multivariate Analysis **176**, 104584 (2020)

[16] Wang, Y., Wang, G., Wang, L., Ogden, R.T.: Simultaneous confidence corridors for mean functions in functional data analysis of imaging data. Biometrics, forthcoming (2020)

Chapter 22
Single Functional Index Model under Responses MAR and Dependent Observations

Nengxiang Ling, Lilei Cheng and Philippe Vieu

Abstract This contribution deals with the estimation of the functional single index regression model (FSIRM) with responses missing at random (MAR) for strong mixing time series data. Some asymptotic properties such as the uniform almost complete convergence rate and asymptotic normality of the estimator are obtained respectively under some general conditions.

22.1 Introduction

The single index model had been employed to reduce the dimensionality of data, and to avoid the "dimension disaster" problem while maintaining the advantages of nonparametric smoothing in multivariate regression case. Furthermore, these ideas have been first extended to the functional setting by [8] for functional regression problems, which leads to the functional single index regression model (FSIRM) below:

$$Y = r(\langle \theta, \chi \rangle) + \varepsilon, \tag{22.1}$$

where $r(\cdot)$ is an unknown real link function from $\mathbb{R}$ to $\mathbb{R}$, $\theta = \{\theta(t), t \in I\}$ is a functional single index set in a separable Hilbert space $\mathcal{H}$ defined on a compact interval I with the inner product $\langle \cdot, \cdot \rangle$ and ε is a random error with $E(\varepsilon \mid \chi) = 0,\ a.s.$ Here, the explanation of Y given $\chi = \chi(t)$ $(t \in I)$ is done through the functional single index θ in $\mathcal{H}$ as $E\big(Y|\chi\big) = E\big(Y| \langle \theta, \chi \rangle\big)$.

Many researchers have paid attention to the functional single index model. For example, [1] proposed to estimate the unknown functional single index via the cross

Nengxiang Ling (✉)
School of Mathematics, Hefei University of Technology, China, e-mail: hfut.lnx@163.com

Lilei Cheng
School of Mathematics, Hefei University of Technology, China, e-mail: hfutcll@163.com

Philippe Vieu
Institut de Mathématiques, Université Paul Sabatier, Toulouse, France,
e-mail: philippe.vieu@math.univ-toulouse.fr

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_22

validation technique. [2] considered the estimation of the conditional density of a scalar response variable Y, given a Hilbertian random variable χ when the observations are linked with a single-index structure. The pointwise and the uniform almost complete convergence rate of the kernel estimate of this model were established. [16] extended the research to the α-mixing case. Meanwhile, [3] obtained some asymptotic results of a nonparametric conditional cumulative distribution estimator with applications in the functional single index model for time series data. [5] investigated a class of functional partially linear single index models, and a profile least squares approach combined with local constant smoothing for estimating the slope function and the link function were proposed in the model. [22] studied a flexible single-index partially functional linear regression model, the convergence rates and asymptotic normality of the estimators were obtained under some mild conditions. [21] investigate the large-sample estimation and inference in multivariate single-index models, and some asymptotic properties of the model are also established.

On the other hand, we notice that all the contributions involved above are in the case of the samples being observed completely. However, in many practical works such as market surveys, medical studies, reliability test and so on, some pairs of observations may be incomplete which is often called missing data. Statistical analysis with missing data is a very difficult task since missing data themselves contain little or no information about missing data mechanism in most cases. The fundamental and most widely used assumption about missing data mechanism is the missing responses at random. We can quote: [4], [18], [19], [20], [12], [6] and references therein for explanatory variables being finite dimensionality. When explanatory variables are infinite case or they are of functional feature, only very few literature was reported to investigate statistical models with missing data. For example, [9] first proposed to estimate the mean of a scalar response based on an i.i.d. functional sample in which the explanatory variables are observed completely and the response variables are missing at random. [15] investigated the asymptotic properties of regression operator estimate for functional stationary ergodic data with missing responses at random. While [14] and [7] study respectively the semi-functional partially linear regression model and the functional linear model with missing scalar responses at random in the case of i. i. d. sample. [11] investigate the inferential procedures for partially observed functional data.

Inspired by all the papers above, our contribution in this paper is to investigate the functional single index regression model (22.1) with missing responses at random in the case of strong mixing functional time series data. For more details in this direction, see [14]. Let χ be observed completely, and $\delta = 0$ if Y is missing, otherwise $\delta = 1$. Furthermore, let the missing mechanism be such that

$$P\,(\delta = 1|\chi, Y) = P(\delta = 1|\chi) = p(\chi).$$

We first recall the definition of strong mixing dependence. A process $\{(\chi_i, Y_i), i \geq 1\}$ is called strong mixing with mixing coefficient $\alpha(n)$, if $\alpha(n) = \sup_\kappa \sup_{A \in \mathcal{A}_1^\kappa} \sup_{B \in \mathcal{A}_{\kappa+n}^{+\infty}} |P(A \cap B) - P(A)P(B)| \longrightarrow 0,\ as\ n \longrightarrow \infty$, where $\mathcal{A}_j^\kappa$ denotes the σ-algebra generated by the random vectors $\{(\chi_i, Y_i), j \leq i \leq k\}$.

The process $\{(\chi_i, Y_i), i \geq 1\}$ is said to be arithmetically strong mixing with order $a > 0$, if $\exists C > 0, \ \alpha(n) \leq Cn^{-\alpha}$.

22.2 Model and Methodology

22.2.1 Modelling with Responses MAR and Estimators

Following model (22.1), let $\{(\chi_i, Y_i, \delta_i), 1 \leq i \leq n\}$ be arithmetically strong mixing functional data with identically distribution as $(\boldsymbol{\chi}, Y, \delta)$, where χ_i, $\boldsymbol{\chi}$ take value in a separable Hilbert space $\mathcal{H}$ with scalar product $\langle \cdot, \cdot \rangle$ and its norm $\| \cdot \| = \langle \cdot, \cdot \rangle^{1/2}$, and Y_i, Y take value in $\mathbb{R}$. That is

$$Y_i = r(\langle \theta, \boldsymbol{\chi}_i \rangle) + \varepsilon_i, \ i = 1, 2, ..., n.$$

with $E(\varepsilon_i \mid \chi_i) = 0, a.s.$ $\delta_i = 0$ if Y_i is missing and $\delta_i = 1$ otherwise. We also assume that $P(\delta_i = 1 | Y_i, \chi_i) = P(\delta_i = 1 | \chi_i) = p(\chi_i), i = 1, 2, ..., n$. Meanwhile, let the semi-metric $d_\theta(\cdot, \cdot)$ with the functional single index θ be defined as $d_\theta(\chi_1, \chi_2) := |\langle \chi_1 - \chi_2, \theta \rangle|$. The kernel estimator of $r(\langle \theta, \chi \rangle)$ is constructed as following:

$$\widehat{r}_n(\theta, \chi) = \frac{\sum_{i=1}^n \delta_i Y_i K(h^{-1} d_\theta(\chi, \chi_i))}{\sum_{i=1}^n \delta_i K(h^{-1} d_\theta(\chi, \chi_i))} := \frac{\widehat{r}_{2n}(\theta, \chi)}{\widehat{r}_{1n}(\theta, \chi)}, \tag{22.2}$$

where $\widehat{r}_{1n}(\theta, \chi) = \frac{1}{n} \sum_{i=1}^n \delta_i \Delta_i(\theta, \chi)$ and $\widehat{r}_{2n}(\theta, \chi) = \frac{1}{n} \sum_{i=1}^n \delta_i Y_i \Delta_i(\theta, \chi)$, with $\Delta_i(\theta, \chi) =: \frac{K(h^{-1} d_\theta(\chi, \chi_i))}{EK(h^{-1} d_\theta(\chi, \chi_i))}$ and $K(\cdot)$ is a kernel functions, $h := h_n > 0$ is a sequence of bandwidths tending to zero as n goes to infinity.

Obviously, the estimator (22.2) depends on the functional single index $\theta = \{\theta(t), t \in I\}$. However, in practice, θ is usually unknown and must be estimated. Here, for any θ, we borrow an idea from [5] by using a profile least squares approach combined with local smoothing constant technique to estimate the functional single index θ. So we need to minimize the following weighted sum:

$$\sum_{i=1}^n \sum_{j:j \neq i} \left(Y_j \delta_j - r(\langle \theta, \chi_i \rangle)\right)^2 K(h^{-1} d_\theta(\chi_j, \chi_i))/h, \tag{22.3}$$

where $K(\cdot)$ is a kernel function and h is a bandwidth with $0 < h = h_n \rightarrow 0$. Let $\overline{\omega}_i$ be the solution of minimizing (22.3) which is presented as

$\overline{\omega}_i = \dfrac{\sum_{j:j \neq i} Y_j \delta_j K_{ij}}{\sum_{j:j \neq i} K_{ij}}$, where $K_{ij} = K(h^{-1} d_\theta(\chi_j, \chi_i))$. Then, we obtain the profile least squares estimator by minimizing the profile least squares function below:

$$S(\theta) = \sum_{i=1}^n \left(Y_i \delta_i - \overline{\omega}_i\right)^2 \tag{22.4}$$

subject to $\langle \theta, \theta \rangle = \|\theta\| = 1$ and $\theta(t_0) = 1$ for $t_0 \in I$. For convenience, similar to [5], we have

$$\theta(t) = \sum_{s=1}^{\infty} \alpha_s \varphi_s(t),\ t \in I, \tag{22.5}$$

where $\varphi_1(t), \varphi_2(t), ...$ are chosen as the principal component bases of $\chi(t)$ constructed from covariance function $K_\chi(u, v) = Cov(\chi(u), \chi(v))$ of the random processes $\chi(t)$. However, the basis functions $\varphi_s(t)$ and α_s are unknown in practice, and need to be estimated in order to obtain estimator of $\theta(t)$. For this purpose, we consider the empirical version of $K_\chi(u, v)$ given by

$$\widehat{K}_\chi(u, v) = \frac{1}{n} \sum_{i=1}^{n} \{\chi_i(u) - \overline{\chi}(u)\}\{\chi_i(v) - \overline{\chi}(v)\},$$

where $\overline{\chi}(v) = \frac{1}{n} \sum_{i=1}^{n} \chi_i(t)$, the eigenfunctions $\widehat{\varphi}_s(t)$ of the covariance operator associated with $K_\chi(u, v)$ is the estimators of the basis functions $\varphi_s(t)$ for $s = 1, 2, ...$ By (22.5) and selecting a suitable positive integer τ, it follows that

$$\theta(t) \approx \sum_{s=1}^{\tau} \alpha_s \widehat{\varphi}_s(t). \tag{22.6}$$

Thus, combining (22.4) with (22.6) leads to

$$S(\alpha_1, ..., \alpha_\tau) = \sum_{i=1}^{n} (Y_i - \overline{\omega}_i)^2, \tag{22.7}$$

subject to $\sum_{s=1}^{\tau} \alpha_s^2 = 1$ and $\sum_{s=1}^{\tau} \alpha_s \widehat{\varphi}_s(t_0) > 0$ (i.e. $\theta(t_0) > 0$). Now, following the same steps as [5], we can find the optimal solutions $\widehat{\alpha}_s$ of (22.7), $s = 1, 2, ...\tau$. Then, the estimator of $\theta(t)$ is obtained by

$$\widehat{\theta}(t) = \sum_{s=1}^{\tau} \widehat{\alpha}_s \widehat{\varphi}_s(t). \tag{22.8}$$

Hence, the estimation $\widehat{\theta}(t)$ of $\theta(t)$ depends on the choice of τ although our main interest lies in the case where the scalar responses is MAR for functional time series data.

22.2.2 Some Notations and Assumptions

Let $S_\mathcal{H}$, $\Theta_\mathcal{H}$ be a compact subset of $\mathcal{H}$, and $N_{1,n}$, $N_{2,n}$ be the minimal number of open balls with radius ε in $\mathcal{H}$ which is necessary to cover $S_\mathcal{H}$ and $\Theta_\mathcal{H}$ with centers $\chi_1, ..., \chi_{N_{1,n}}$ and $\theta_1, ..., \theta_{N_{2,n}}$ respectively. For $i = 1, 2..., N_{1,n}$, denote $K_i(\theta, \chi) := K(h^{-1}|\langle \chi - \chi_i, \theta \rangle|)$ and $B_\theta(\chi, h) =: \{\mathcal{Y} \in \mathcal{H}| \langle \chi - \mathcal{Y}, \theta \rangle \leq h\}$ for $\forall \chi \in \mathcal{H}$. Let $\Psi_{S_\mathcal{H}}(\varepsilon) = \log(N_{1,n})$ and $\Psi_{\Theta_\mathcal{H}}(\varepsilon) = \log(N_{2,n})$ be the Kolmogrov's ε-entropy of $S_\mathcal{H}$ and $\Theta_\mathcal{H}$ respectively. For a fixed $\chi \in S_\mathcal{H}$ and $\theta \in \Theta_\mathcal{H}$, denote $t(\chi) = \arg\min_{t \in \{1,2,...,N_{1,n}\}} \|\chi - \chi_t\|$ and $k(\theta) = \arg\min_{k \in \{1,2,...,N_{2,n}\}} \|\theta - \theta_k\|$ respectively.

The $\widehat{\theta}(t)$ depends on the suitable choice of positive integer τ which the parametric can be obtained by cross-validation method, in generally, the maximum value of τ no more than 10. The influence of the mixing structure on the rates of convergence will be shown through the quantities:
$s_{n,1}^2 = \sum_{i=1}^n \sum_{j=1}^n |Cov(Y_i \Delta_i(\theta_{k(\theta)}, \chi_{t(\chi)}), Y_j \Delta_j(\theta_{k(\theta)}, \chi_{t(\chi)}))|$,
$s_{n,2}^2 = \sum_{i=1}^n \sum_{j=1}^n |Cov(Y_i \nabla_i, Y_j \nabla_j)|$,
$s_{n,3}^2 = \sum_{i=1}^n \sum_{j=1}^n |Cov(Y_i \Gamma_i, Y_j \Gamma_j)|$,
$s_{n,4}^2 = \sum_{i=1}^n \sum_{j=1}^n |Cov(\nabla_i, \nabla_j)|$,
$s_{n,5}^2 = \sum_{i=1}^n \sum_{j=1}^n |Cov(\Gamma_i, \Gamma_j)|$,
$s_{n,6}^2 = \sum_{i=1}^n \sum_{j=1}^n |Cov(\Delta_i(\theta_{k(\theta)}, \chi_{t(\chi)}), \Delta_j(\theta_{k(\theta)}, \chi_{t(\chi)}))|$, where
$\nabla_i =: \frac{1}{\phi(h)} I_{\{B_\theta(\chi,h) \bigcup B_\theta(\chi_{t(\chi)},h)\}}(\chi_i)$,
$\Gamma_i =: \frac{1}{\phi(h)} I_{\{B_\theta(\chi_{t(\chi)},h) \bigcup B_{k(\theta)}(\chi_{t(\chi)},h)\}}(\chi_i)$ and $I_A(\cdot)$ is an indicative function of a set A.

Denote $s_n^{\prime 2} = \max\left\{s_{n,1}^2, s_{n,2}^2, s_{n,3}^2\right\}$, $s_n^{\prime\prime 2} = \max\left\{s_{n,4}^2, s_{n,5}^2, s_{n,6}^2\right\}$. Throughout this contribution, let C, C_1 and $C_2, \ldots$ be some positive constants not depending on n, which may take different values in each appearance.

(A1) There exists a differentiable function $\phi(\cdot)$ such that, for $\forall \chi \in S_{\mathcal{H}}$, $\forall \theta \in \Theta_{\mathcal{H}}$, $0 < C_1\phi(h) \leq P\big(\chi \in B_\theta(\chi, h)\big) =: \phi_{\theta,\chi}(h) \leq C_2\phi(h) < \infty$ and $\exists \xi_0 > 0, \forall \xi < \xi_0, \phi^{'}(\xi) < C$.

(A2) The kernel $K(\cdot)$ is a positive bounded function supported on $[0, 1]$ and is differentiable on $[0, 1]$ such that, $\exists C_1, C_2, -\infty < C_1 < K^{'}(t) < C_2 < 0$ for $0 < t < 1$.

(A3) $r(\langle\theta, .\rangle)$ is such that: $\exists \beta > 0, \forall (\chi_1, \chi_2) \in S_{\mathcal{H}} \times S_{\mathcal{H}}$,
$|r(\langle\theta, \chi_1\rangle) - r(\langle\theta, \chi_2\rangle)| \leq C d_\theta(\chi_1, \chi_2)^\beta$.

(A4) There exist $a > 1$ and $b > 2$ such that $s_{n,j}^{-(a+1)} = o(n^{-b})$ for $j = 1, \ldots, 6$.

(A5) For n being large enough, the Kolmogrov's ε-entropy of $S_{\mathcal{H}}$ and $\Theta_{\mathcal{H}}$ for $\varepsilon = \frac{\log n}{n}$ satisfies:

(i) $$\frac{(\log n)^2}{n\phi(h)} < \Psi_{S_{\mathcal{H}}}\left(\frac{\log n}{n}\right) + \Psi_{\Theta_{\mathcal{H}}}\left(\frac{\log n}{n}\right) < \frac{n\phi(h)}{\log n},$$

(ii) $\sum_{n=1}^\infty (N_n)^{1-\gamma} < \infty$, for some $\gamma > 1$ and $N_n = N_{1,n}.N_{2,n}$,

(iii) $n\phi(h) = O((\log n)^2)$.

(A6)

(i) $\exists \nu > 2$, such that $E|Y|^\nu < \infty$. For a fixed $\theta \in S_{\mathcal{H}}$ and $\forall u \in S_{\mathcal{H}}$.

(ii) $E[(Y - r(\langle\theta, \chi\rangle))^2 | \chi = u] = g_2(u, \theta)$ and $p(\cdot)$ are continuous in a neighborhood of χ respectively, that is
$\sup_{\{u: d_\theta(\chi,u) \leqslant h\}} |g_2(u, \theta) - g_2(\chi, \theta)| = o(1),$ as $h \to 0$,
$\sup_{\{u: d_\theta(\chi,u) \leqslant h\}} |p(u) - p(\chi)| = o(1),$ as $h \to 0$.

(iii) Let $g_\nu(u, \theta) = E\big[\big|Y - r(\langle\theta, \chi\rangle|^\nu \big| \chi = u\big]$ be continuous in some neighborhood of χ.

(A7)

(i) $0 < \sup_{i \neq j} P[(\chi_i, \chi_j) \in B_\theta(\chi, h) \times B_\theta(\chi, h)] \leqslant \psi_{\theta,\chi}(h)$, where $\psi_{\theta,\chi}(h) \to 0$ as $h \to 0$ with $\frac{\psi_{\theta,\chi}(h)}{\phi^2_{\theta,\chi}(h)} = O(1)$.

(ii) There exists an function $\tau_{\theta,\chi}(\cdot)$ satisfying: $\forall t \in [0,1]$, $\lim\limits_{n\to\infty} \frac{\phi_{\theta,\chi}(ht)}{\phi_{\theta,\chi}(h)} = \tau_{\theta,\chi}(t)$.

(A8) The bandwidths $h = h_n$ satisfies:

(i) $n\phi_{\theta,\chi}(h) \to \infty$, as $n \to \infty$.

(ii) $nh^{2\beta}\phi_{\theta,\chi}(h) \to 0$, as $n \to \infty$.

(A9) For some $\nu > 2$ and $m > 1 - 2/\nu$ such that $\sum_{n=1}^{\infty} n^m [\alpha(n)]^{1-2/\nu} < \infty$.

Comments on the assumptions: Assumption (A1) characterizes the concentration of the explanatory variable in small balls. Similar to the discussions in [10], (A2) and (A3) are the quite usual conditions on the kernel function for nonparametric FDA. (A4) shows the covariance structure of the dependent sample, see, for instance, [17] and [10] respectively for details; (A5) is used to obtain the uniform consistency rate. (A6)(ii) and (A6)(iii) stand as local continuous conditions, which is necessary to establish the main results and make the results more concise in this paper. (A7)(i) gives the behavior of the joint distribution of the couple (χ_i, χ_j) in relation to its margin, and also permits us to present an explicitly asymptotic variance term, respectively. (A8)(ii) will be used to remove the bias term in the asymptotic normality results. (A9) is a standard assumption on the strong mixing coefficient.

22.3 Asymptotic Properties

Theorem 1 *Under assumptions (A1)–(A5), we have that, as $n \to \infty$,*

$$\sup_{\theta \in \Theta_{\mathcal{H}}} \sup_{\chi \in S_{\mathcal{H}}} |\widehat{r}_n(\theta, \chi) - r(\langle \theta, \chi \rangle)| = \\ = O(h^\beta) + O_{a.co}\left(\sqrt{\frac{s_n^{*2}\left(\Psi_{S_{\mathcal{H}}}(\log n / n) + \Psi_{\Theta_{\mathcal{H}}}(\log n / n)\right)}{n^2}}\right), \tag{22.9}$$

where $s_n^{*2} = max\left\{s_n'^2, s_n''^2\right\}$.

Theorem 2 *Under assumptions (A1)–(A3) and (A6)–(A9), then we get that,*

$$(n\phi_{\theta,\chi}(h))^{1/2}\left(\widehat{r}_n(\theta,\chi) - r(\langle\theta,\chi\rangle)\right) \xrightarrow{D} N(0, \sigma^2(\theta,\chi)), \ as\ n \to \infty, \tag{22.10}$$

where $\theta \in \{\theta(t),\ t \in I\}$ and $\sigma^2(\theta,\chi) := \frac{M_2}{M_1^2}\frac{g_2(\chi,\theta)}{p(\chi)}$

with $M_j = K^j(1) - \int_0^1 (K^j)'(u)\tau_{\theta,\chi}(u)du$, for $j = 1, 2$, and $\xrightarrow{D}$ means the convergence in distribution.

22.4 Conclusion

This contribution investigates the estimation of the functional single index regression model (FSIRM) with responses missing at random (MAR) for strong mixing time series data. The large sample properties such as the uniform almost complete convergence rate and asymptotic normality of the estimator are obtained respectively under some general conditions. What is presented here is mainly issued from [13] in which the detail proofs and extensive simulation studies as well as deeper real data analysis of the method can be found.

References

[1] Ait-Saïdi, A., Ferraty, F., Kassa, R., Vieu, P.: Cross-validated estimation in the single functional index model.Statistics **42**(6), 475–494 (2008)

[2] Attaoui, S., Laksaci, A., Ould-Said, E.: A note on the conditional density estimate in the single functional index model. Statistics and Probability Letters **81**(1), 45–53 (2011)

[3] Attaoui, S., Ling, N.X.: Asymptotic results of a nonparametric conditional cumulative distribution estimator in the single functional index modeling for time series data with applications. Metrika **79**, 485–511 (2016)

[4] Cheng, P.E.: Nonparametric Estimation of Mean Functionals With Data Missing at Random. Journal of the American Statistical Association **89**, 81–87 (1994)

[5] Ding, H., Liu, Y.H., Xu, W.C., Zhang, R.Q.: A class of functional partially linear single-index models. Journal of Multivariate Analysis **161**, 68–82 (2017)

[6] Efromovich, S.: Nonparametric regression with predictors missing at random. Journal of the American Statistical Association **106**, 306–319 (2011)

[7] Febrero-Bande, M., Galeano, P., Gonzalez-Manteiga, W.: Estimation imputation and prediction for the functional linear model with scalar response with responses missing at random. Computational Statistics and Data Analysis **131**, 91–103 (2019)

[8] Ferraty, F., Peuch, A., Vieu, P.: Modèle à indice fonctionnel simple. C. R. Math. Acad. Sci. Paris **336**(12), 1025–1028 (2003)

[9] Ferraty, F., Sued, M., Vieu, P.: Mean estimation with data missing at random for functional covariables. Statistics **47**(4), 688–706 (2013)

[10] Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis, Theory and Practice. Springer (2006)

[11] Kraus, D.: Inferential procedures for partially observed functional data. Journal of Multivariate Analysis **173**, 583–603 (2019)

[12] Liang, H., Wang, S., Carroll, R.: Partially linear models with missing response variables and error-prone covariates. Biometrika **94**, 185–198 (2007)

[13] Ling, N.X., Cheng, L.L., Vieu, P.: Missing responses at random in functional single index model for time series data. Submitted to Statistical Papers, revised (2020)

[14] Ling, N.X., Kan, R.: Semi-functional partially linear regression model with responses missing at random. Metrika, **82**(1), 39–70 (2019)

[15] Ling, N.X., Liang, L.L., Vieu, P.: Nonparametric regression estimation for functional stationary ergodic data with missing at random. Journal of Statistical Planning and Inference **162**, 75–87 (2015)

[16] Ling, N.X., Li, Z.H.: Conditional Density Estimation in the Single Functional Index Model for α-Mixing Functional Data. Communications in Statistics Theory and Methods **43**, 441–454 (2014)

[17] Ling, N.X., Xu, Q.: Asymptotic normality of conditional density estimation in the single index model for functional time series data. Statistics and Probability Letters **82**, 2235–2243 (2012)

[18] Little, R., Rubin, D.: Statistical Analysis with Missing Data. Second Edition.Wiley, New York (2002)

[19] Nittner, T.: Missing at random (MAR) in nonparametric regression - A simulation experiment. Statistical Methods and Applications **12**(2), 195–210 (2003)

[20] Tsiatis, A.: Semiparametric Theory and Missing Data. Springer, New York (2006)

[21] Wu, J.W., Peng, H.X., Tu, W.Z.: Large-sample estimation and inference in multivariate single-index models. Journal of Multivariate Analysis **171**, 382–396 (2019)

[22] Yu, P., Du, J., Zhang, Z.: Single-index partially functional linear regression model, Statistical Papers, 1–17 (2018)

Chapter 23
O2S2 for the Geodata Deluge

Alessandra Menafoglio, Davide Pigoli and Piercesare Secchi

Abstract We illustrate a few recent ideas of Object Oriented Spatial Statistics (O2S2), focusing on the problem of kriging prediction in situations where a global second order stationarity assumption for the random field generating the data is not justifiable or the space domain of the field is complex. By localizing the analysis through the Random Domain Decomposition algorithm, we build ensembles of local predictors eventually aggregated in an ultimate one. The localization allowed by the algorithm is also effective for dealing with data which are mildly non-Euclidean and can be locally linearized, as it happens for data embedded in a Riemannian manifold.

23.1 Introduction: Object Oriented Data Analysis for Spatially Dependent Complex Data

Public and private companies, national statistical offices, healthcare organizations and systems, space agencies, research laboratories and universities collect and store an ever increasing amount of data which are referenced in space or time. Data are captured by infrastructures distributed in space, like mobile phone networks tracking people and objects moving over an urban maze or sensor systems recording over time the composition of chemical species measured in sampled locations of an estuarine region or in a large-scale ground-water body. Additive manufacturing, which is feeding the 4th industrial revolution, requires real-time monitoring of parts

Alessandra Menafoglio
MOX - Department of Mathematics, Politecnico di Milano, Milano, Italy,
e-mail: alessandra.menafoglio@polimi.it

Davide Pigoli
Department of Mathematics, King’s College London, London, United Kingdom,
e-mail: davide.pigoli@kcl.ac.uk

Piercesare Secchi (✉)
MOX - Department of Mathematics, Politecnico di Milano, Milano, Italy and Center for Analysis Decisions and Society, Human Technopole, Milano, Italy, e-mail: piercesare.secchi@polimi.it

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_23

represented as shapes and identified as space dependent data points of a manifold. Steadily more, these data are high dimensional and complex – like functions, tensors, graphs and networks – due to the advance of a new generation of sensors and diagnostic devices able to measure reality in a pulverized format.

The availability of massive geodata is pushing the demand for new statistical paradigms, oriented by the goals of the analysis and the nature of the data the analysis is based upon. Following the Object Oriented Data Analysis approach first advocated by Wang and Marron [17], the atoms of the statistical inquiry (e.g., vectors, curves, operators, networks) are indivisible objects which should be modeled as points of a mathematical space whose dimensionality, topology and geometrical properties must not injure the data complexity in the face of the goals of the analysis. In this short paper we will pursue an object oriented approach for spatial statistics – christened Object Oriented Spatial Statistics (O2S2) in [7].

Typical problems of O2S2 are those of prediction, classification, regression, data fusion and dimensional reduction; for a recent review, we refer to [10], beside the already mentioned [7]. In the following pages we illustrate some recent advances of the O2S2 system of ideas, leveraging the papers [8, 9] and considering the problem of kriging prediction in situations where the random field generating the object data cannot be assumed to be globally stationary, or the spatial domain supporting the random field is complex, although locally Euclidean. To attack these issues we will illustrate a Random Domain Decomposition (RDD) approach which builds an ensemble of local models to generate predictions eventually aggregated into a final global one. Interestingly, the localization idea RDD is based upon is also instrumental to tackle the kriging problem when the object data are mildly non-Euclidean and can be locally linearized, as it happens when they are elements of a Riemannian manifold.

In the next section we will briefly introduce the notion of kriging in a O2S2 perspective. The following section will sum up the RDD algorithm. Next we will illustrate kriging prediction by means of RDD by summarizing two case studies directed at the environmental monitoring of the Chesapeake Bay. A section with conclusions will close the paper.

23.2 A Gentle Initiation to the O2S2 Perspective on Kriging

Let $x_{s_1}, ..., x_{s_n}$ be n data objects observed in the sampled locations $s_1, ..., s_n$ of a spatial domain D. Assuming that the x_{s_i}'s are point observations of a random field defined on D, the problem is to predict the realization x_{s_0} of the field at an unobserved location s_0 in D. Approaching the problem, we are immediately confronted with two critical issues. The first is related with the complexity of the objects x_{s_i}'s, the second with the complexity of the spatial domain D.

Following the object oriented approach, the random data cloud $\{x_{s_i} : s_0, s_1, ..., s_n \in D\}$ should be embedded in a mathematical space $\mathcal{M}$ – often called the *feature space* – whose geometry must properly account for the nature of the objects and the scope of the analysis. For instance, if the goal is to elicit a representation of the predictor of x_{s_0} in a linear form

$$\sum_{i=1}^{n} \lambda_i x_{s_i},$$

then $\mathcal{M}$ should be linear. Moreover, if we aim at an ordinary kriging predictor [2] where the weights $\lambda_1, ..., \lambda_n$ are optimal in the sense of minimizing

$$E\left[\left\| x_{s_0} - \sum_{i=1}^{n} \lambda_i x_{s_i} \right\|^2\right] \quad \text{subject to} \quad E\left[\sum_{i=1}^{n} \lambda_i x_{s_i}\right] = E[x_{s_0}], \tag{23.1}$$

then $\mathcal{M}$ should be normed – allowing to define the expected value $E[\cdot]$ à la Fréchet. In this case, we also need to be able to measure stochastic dependence between random elements with values in $\mathcal{M}$, something which is conveniently achieved if $\mathcal{M}$ is endowed with a notion of inner product and thus a mean to measure angles.

Kriging prediction when the feature space $\mathcal{M}$ is Hilbert has been extensively treated in the literature: see the review paper [7] and references therein. Notably, in Functional Data Analysis a typical choice for $\mathcal{M}$ is the space L^2 of square integrable functions. Nonetheless, when point-wise and differential information embedded within the functions are both relevant for the analysis, a Sobolev space might be a more appropriate feature space. It might happen that the data objects are constrained functional data, for instance when they can be represented as probability densities, i.e. non negative functions integrating to a positive constant; these *compositional* data can be suitably embedded in a Hilbert space endowed with the generalized Aitchison's geometry (a.k.a., Bayes Hilbert space, [5]). In fact, we will use this specific feature space to maximum advantage in a case study illustrated in Section 23.4 which explores the random field of the densities of oxygen dissolved in the waters of the Chesapeake Bay.

Things do not proceed so smoothly when object data are not Hilbert, for instance when they are points of a Riemannian manifold, a non-linear space which can be approximated by a Hilbert space only locally. While there have been some attempts to introduce new concepts of stochastic dependence in this kind of spaces (see, e.g, [11, 4]), a rigorous treatment of kriging prediction is still out of reach due to the nonlinearity of the Fréchet expectation. To overcome this issue, in [12], we introduced a tangent space model for kriging Riemannian data which grounds on the idea of (i) linearizing the data by projecting them into the tangent space (which is Hilbert), (ii) perform kriging in the tangent space, and (iii) finally transform the prediction back into the original Riemannian feature space. This approach works very well when the variability of the data is not large, i.e. when the data reside in a local neighbor of the feature space and therefore the linear approximation allowed by their projection on the tangent space is good. Local linearization of Riemannian data will be the key point of a case study illustrated in Section 23.4 where we analyze the random field of covariance matrices of dissolved oxygen and water temperature over the Chesapeake Bay. Localization will be obtained by means of the RDD scheme reported in Section 23.3.

A second critical element of complexity may arise when the spatial domain D has a complex topology and is not convex due to the presence of holes or boundaries, or when the appropriate notion of closeness is not captured by the Euclidean distance. These issues make it difficult to leverage Tobler's first law of geography which states that "everything is related to everything else, but near things are more related than distant things" [15]. The problem of analyzing spatial data when their domain of observation has a complex topology has been considered in the functional data literature: for instance, see [13] and references therein. We approach the problem through localization, assuming that the space domain D can be locally approximated by simple Euclidean subdomains. Localization is once again obtained by means of the RDD scheme, which is by now time to put forward.

23.3 Localization through Random Domain Decomposition

The classic workhorse for solving the kriging prediction optimality problem formalized in (23.1) is the semi-variogram [2], whose counterpart when data are Hilbert is the trace semi-variogram [6], defined, for any two locations $s_i, s_j \in D$, as

$$\gamma(s_i, s_j) = \frac{1}{2}\left\{E\left[\|x_{s_i} - x_{s_j}\|^2\right] - \|E\left[x_{s_i}\right] - E\left[x_{s_j}\right]\|^2\right\}.$$

If the random field defined on D and generating the x_{s_i}'s has spatially constant mean and is second-order stationary, then γ can be estimated by fitting a parametric valid model to the empirical trace semi-variogram. This course of action can be extended to the case of mildly non-stationary random fields, where the mean of the field is not constant but can be represented by a linear model, and the residuals are second order stationary: for details, see [6].

For stronger types of non-stationarity and object data, the literature is rather scanty and no all-encompassing procedure is known to us, even for cases where the spatial domain D is a rectangular subset of a Euclidean space. The problem is further complicated by the fact that non-stationarity of the field generating the data often happens in co-occurence with a spatial domain whose topology is complex, with holes or barriers, and the distance measuring closeness between locations is not Euclidean, as it happens, for instance, between locations of an estuarine system lying on opposite sides of a promontory. However, the degree of non-stationarity of the field, as well as the degree of complexity of the domain, may depend on the spatial scale of observation. Indeed, even though the field may appear non-stationary at a global spatial scale, stationarity may be a viable assumption at a local scale. Similarly, a spatial domain which is complex at a global scale, might be locally approximated through a simple Euclidean domain.

For tackling the kriging prediction problem in these situations, the Random Domain Decomposition (RDD) approach was proposed in [8]; its goal is to localize the analysis by means of a *divide et impera* strategy, and generate an ensemble of locally optimal predictors, which are eventually aggregated in a final one. RDD is rooted in the *bagging* idea of Breiman [1] and has a direct predecessor called Bagging Voronoi Algorithm [14]. Indeed, RDD is composed of two stages: a *bootstrap* stage and an *aggregating* stage. The bootstrap stage consists of B iterations: at each iteration the spatial domain D is randomly partitioned in local neighbors, where the notion of neighbor must be consistent with the specific topology characterizing the complexity of D. In each neighbor, a local analysis is performed, assuming stationarity of the field generating the data and Euclidean approximation of the spatial domain. Hence, at a given location $s_0 \in D$, the bootstrap stage results in a set of B predictions, each of them locally optimal conditionally on the realization of the random partition of D. In the final aggregating stage, the B predictions resulting from the bootstrap stage are aggregated in a ultimate prediction, typically by taking a (weighted) average. For further details, we refer to [8].

Consistently with Tobler's law [15], by localizing in space spatially dependent data, RDD has the further effect of localizing data in their feature space. This makes RDD suitable for the analysis of non-Euclidean spatially dependent object data which could be efficiently locally linearized, as is the case when the feature space is a Riemannian manifold. This idea was exploited in [9], using to maximum advantage the tangent space model for kriging Riemannian data introduced in [12], a model which is indeed appropriate when the data variability is not large.

23.4 Monitoring Dissolved Oxygen in the Waters of the Chesapeake Bay

The Chesapeake Bay is the largest estuarine system in USA. The Bay has been monitored for years, to assess the impact of human activities on aquatic variables deemed critical for its ecosystem. Of primary importance is the dissolved oxygen (DO) in the waters of the Bay, since DO is necessary for the life of most marine species. The areas of the Bay where DO is below 2 mg/l are called *Dead zones*. In these areas most of the marine species suffocate.

Figure 23.1b shows the outline of the Chesapeake Bay and the locations of 110 monitoring stations (dots). In each of these stations, 17 measurements of DO were recorded in the time period 1990-2006 [source: US Environmental Protection Agency Chesapeake Bay Program (US EPA-CBP)]. No significant autocorrelation exists, along the years, for the time series of DO. In the figure, each dot indicating the position of a monitoring station has been colored according to the corresponding value of the DO sample mean. The same color is used in the graph in Figure 23.1a, where the smoothed probability density functions (pdf) of DO are represented: these pdf's are the object data of this first case study, detailed in [8], whose primary goal is to predict the DO pdf at the unobserved locations of the Bay.

We embed the DO pdf's in the Bayes Hilbert feature space introduced in [16]. This is the space $\mathcal{B}^2(I)$ of real valued positive functions defined on an interval $I \subset R$ of length η, whose logarithm is squared-integrable. The space $\mathcal{B}^2(I)$ is endowed with the equivalence relation of proportionality, and is equipped with a separable Hilbert structure if, for any $f, g \in \mathcal{B}^2(I)$ and $\alpha \in R$, the following operations

$$(f \oplus g)(t) = \frac{f(t)g(t)}{\int_I f(s)g(s)\,\mathrm{d}s}, \quad (\alpha \odot f)(t) = \frac{f(t)^\alpha}{\int_I f(s)^\alpha\,\mathrm{d}s}, \quad t \in I,$$

and inner product,

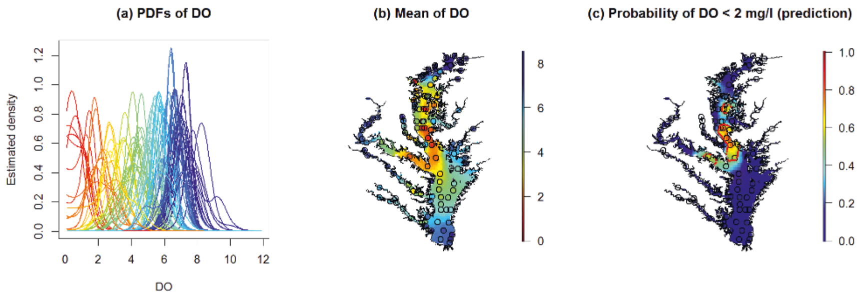


Fig. 23.1 Distribution of DO in the Chesapeake Bay. Prediction in panels (b) and (c) are obtained by using 16 local neighbors in each of the 100 bootstrap iterations.

$$\langle f, g \rangle = \frac{1}{2\eta} \int_I \int_I \ln \frac{f(t)}{f(s)} \ln \frac{g(t)}{g(s)} \, \mathrm{d}t \, \mathrm{d}s,$$

are defined. See [16] and references therein for more details.

Next come the realization that the random field generating the DO pdf's is not stationary and the spatial domain D of the estuarine system, i.e. the area of the Bay covered by water, is non convex, with irregular boundaries, promontories and islands. Accordingly, the distance capturing the notion of closeness in the Bay is non Euclidean, since even if a short air distance separates two points lying on opposite sides of a promontory, the land separating them represents a barrier for the distribution of DO (i.e. large water distance). Hence, for computing kriging predictions of DO pdf's, we follow the RDD, having represented the Bay through a constrained Delaunay triangulation (for details see [8]). In the bootstrap stage of RDD, local neighbors of D are obtained based on the graph-based metric implied by the triangulation. Aggregation of the local predictors is secured by simple averaging in $\mathcal{B}^2(I)$. Details for setting the model parameters for this specific RDD implementation are given in [8]. Note that kriging the entire pdf's, instead of, e.g., their summary statistics (mean, variance, or selected quantiles), allows projecting the full information content embedded in pdf's to unsampled locations of D. For instance, Figure 23.1b-c show, respectively, the mean of the predicted pdf's, and the probability of $DO < 2mg/l$: dead zones are identified by a contour line representing the value of this probability being equal to 1/2.

Dissolved oxygen in water is influenced by the water temperature (WT). Hence it becomes of interest to study the spatial variation of the covariance matrix of DO and WT over the Chesapeake Bay. The dataset we consider for this second case study consists of the sample covariance matrices of DO and WT, estimated at 144 locations within the Bay where more than 10 joint measurements, taken along the period 1990-2006, were available [source: US Environmental Protection Agency Chesapeake Bay Program (US EPA-CBP)]. The object data are now covariance matrices and they are embedded in the feature space $PD(2)$ of the 2x2, symmetric, positive definite matrices with real entries. The goal of the analysis is to predict the covariance matrix between DO and WT at an unobserved location of the Bay; details of this case study are reported in [9].

The feature space $PD(2)$ is a convex cone, subset of the Hilbert space $Sym(p)$ of the 2x2, symmetric matrices with real entries. Although $PD(2)$ is not a linear space, it can be endowed with a metric [3]. A natural choice is the invariant under affine transformation metric defined, for all $\Psi_1, \Psi_2 \in PD(2)$, as

$$d_R(\Psi_1, \Psi_2) = || \log(\Psi_1^{\frac{1}{2}} \Psi_2 \Psi_1^{\frac{1}{2}}) ||,$$

where $\log(\cdot)$ is the logarithm matrix, and $|| \cdot ||$ is the norm defined on $Sym(p)$ when this space is endowed with the Frobenius inner product $\langle A_1, A_2 \rangle = \text{trace}(A_1^T A_2)$, for $A_1, A_2 \in Sym(p)$. For any $\Psi \in PD(2)$, geodesics in $PD(2)$ passing through Ψ, as well as associated exponential and logarithmic maps – i.e., maps to and from the linear tangent spaces in Ψ , identified with $Sym(p)$, – are defined consistently with d_R; for references and details we refer to [9].

Applying the RDD scheme along the same line described above for predicting the pdf's of DO, we are able to localize the kriging prediction to neighbors of D where the variability of the objects is not large and therefore the tangent space model for kriging covariances, introduced in [9], is reasonable. Results are shown in Figure 23.2 which represents, for different locations of the Bay, the sd's of DO and WT, together with their correlations, simultaneously estimated as elements of the corresponding covariance matrix.

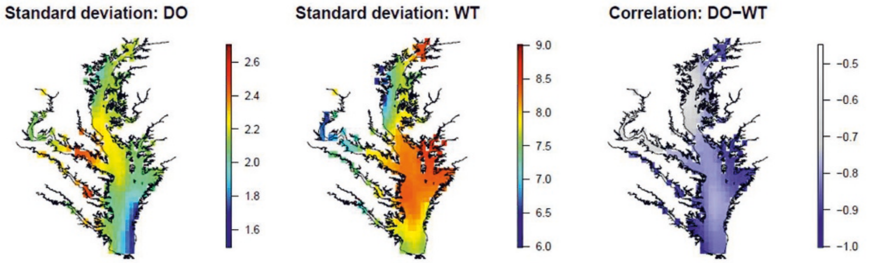


Fig. 23.2 Predictions of covariance matrices between DO and WT in the Chesapeake Bay, using 10 local neighbors in each of the 100 bootstrap iterations.

23.5 Conclusions

O2S2 is a system of ideas which meets the growing demand for statical models and algorithms able to drive the analyses required by the geodata deluge. In this short illustrative paper, we pointed to our main contributions to the problem of kriging object data, when the assumption of stationarity of the random field generating them is not tenable, or the spatial domain is complex.

References

[1] Breiman, L.: Bagging predictors. Machine Learning **24**, 123–140 (1996)

[2] Cressie, N.: Statistics for Spatial data. John Wiley & Sons, New York (1993)

[3] Dryden I.L., Koloydenko A., Zhou D.: Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. The Annals of Applied Statistics **3**(3), 1102–1123 (2009)

[4] Dubey, P., Müller, H.G.: Functional models for time-varying random objects. arXiv:1907.10829 (2019)

[5] Egozcue, J., Díaz-Barrero, J.L., Pawlowsky-Glahn, V.: Hilbert space of probability density functions based on Aitchison geometry. Acta Mathematica Sinica, English Series **22**(4), 1175–1182 (2006)

[6] Menafoglio, A., Secchi, P., Dalla Rosa, M.: A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. Electronic Journal of Statistics **7**, 2209–2240 (2013)

[7] Menafoglio, A., Secchi, P.: Statistical analysis of complex and spatially dependent data: A review of object oriented spatial statistics. European Journal of Operational Research **258**(2), 401– 410 (2017)

[8] Menafoglio, A., Gaetani, G., Secchi, P.: Random domain decompositions for object-oriented kriging over complex domains. Stochastic Environmental Research and Risk Assessment **32**, 3421–3437 (2018)

[9] Menafoglio, A., Pigoli, D., Secchi, P.: Kriging Riemannian Data via Random Domain Decompositions. MOX - Report No. 64/2018, Dipatimento di Matematica, Politecnico di Milano (2018)

[10] Menafoglio, A., Secchi, P.: O2S2: A new venue for computational geostatistics. Applied Computing and Geosciences **2**, 100007 (2019)

[11] Pigoli, D., Secchi, P.: Estimation of the mean for spatially dependent data belonging to a Riemannian manifold. Electronic Journal of Statistics **6**, 1926–1942 (2012)

[12] Pigoli, D., Menafoglio, A., Secchi, P.: Kriging prediction for manifold-valued random field. J. Multivar. Anal. **145**, 117–131 (2016)

[13] Sangalli, L.M., Ramsay, J.O., Ramsay, T.O.: Spatial spline regression models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **75**(4), 681–703 (2013)

[14] Secchi, P., Vantini, S., Vitelli, V.: Bagging Voronoi classifiers for clustering spatial functional data. International Journal of Applied Earth Observation and Geoinformation **22**, 53–64 (2013)

[15] Tobler, W.: A computer movie simulating urban growth in the Detroit region. Economic Geography **46**, 234–240 (1970)

[16] van den Boogaart, K.G., Egozcue, J.J., Pawlowsky-Glahn V.: Bayes Hilbert spaces. Aust N Z J Stat **56**, 171–194 (2014)

[17] Wang, H., Marron, J.S.: Object oriented data analysis: sets of trees. Annals of Statistics **35**(5), 1849–1873 (2007)

Chapter 24
Riemannian Distances between Covariance Operators and Gaussian Processes

Minh Hà Quang

Abstract In this work we study several recently formulated Riemannian distances between infinite-dimensional positive definite Hilbert-Schmidt operators in the context of covariance operators associated with functional random processes. Specifically, we focus on the affine-invariant Riemannian and Log-Hilbert-Schmidt distances and the family of Alpha Procrustes distances, which include both the Bures-Wasserstein and Log-Hilbert-Schmidt distances as special cases. In particular, we present finite-dimensional approximations of the infinite-dimensional distances and show their convergence to the exact distances. The theoretical formulation is illustrated with numerical experiments on covariance operators of Gaussian processes.

24.1 Introduction

The study of functional data has received increasing interests recently, see e.g. [25, 6, 10]. One particular approach for analyzing functional data has been via the analysis of covariance operators. Recent work along this direction includes [21, 7], which utilize the Hilbert-Schmidt distance between covariance operators and [24, 13], which utilize non-Euclidean distances, in particular the Procrustes distance, also known as Bures-Wasserstein distance. The latter distance corresponds to precisely the $\mathcal{L}^2$-Wasserstein distance between two zero-mean Gaussian measures on Hilbert space in the context of optimal transport and can better capture the intrinsic geometry of the set of covariance operators. In this work, we study other non-Euclidean distances between covariance operators that arise from the Riemannian geometric viewpoint of positive definite Hilbert-Schmidt operators, including in particular the affine-invariant Riemannian and Log-Hilbert-Schmidt distances. [1]

Minh Hà Quang (✉)
RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, 15F, Chuo-kuo, Tokyo, 103-0027, Japan, e-mail: minh.haquang@riken.jp

[1] Many more theoretical results, along with further numerical experiments, will be presented in the longer version of the current work.

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_24

24.2 Covariance Operators Associated with Random Processes

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $I \subset \mathbb{R}$ be compact. Let X be a random process taking values in $\mathcal{L}^2(I)$, that is $X(t) \in \mathcal{L}^2(I)$ $\forall t \in I$, where $\mathcal{L}^2(I)$ is the space of square integrable functions on I under the Lebesgue measure. Assume further that $\mathbb{E}(||X(t)||^2_{\mathcal{L}^2(I)}) < \infty$. The *mean function* associated with X is defined by $m_X(t) = \mathbb{E}(X(t))$. The *covariance function* associated with X is defined by

$$\mathrm{cov}_X(t_1, t_2) = \mathbb{E}([X(t_1) - m_X(t_1))][X(t_2) - m_X(t_2))]). \tag{24.1}$$

Then $\mathrm{cov}_X : I \times I \to \mathbb{R}$ is a positive definite kernel and gives rise to the covariance operator $C_X : \mathcal{L}^2(I) \to \mathcal{L}^2(I)$, which is the integral operator defined by

$$(C_X g)(t) = \int_I \mathrm{cov}_X(t, t')g(t')dt', \quad g \in \mathcal{L}^2(I). \tag{24.2}$$

Let us assume further that cov_X is continuous. Then C_X is a positive, self-adjoint, trace class operators on $\mathcal{L}^2(I)$. In the following, we study different distances, particularly Riemannian distances, between two covariance operators C_X and C_Y associated with two random processes X and Y, respectively.

24.3 Finite-dimensional Distances

Let A, B be two covariance matrices corresponding to two Borel probability measures in $\mathbb{R}^n$, then $A, B \in \mathrm{Sym}^+(n)$, the set of $n \times n$ real, symmetric, positive semi-definite matrices. Examples of distance functions that have been studied on $\mathrm{Sym}^+(n)$ include

1. Euclidean (Frobenius) distance $d_E(A, B) = ||A - B||_F$, where $||\ ||_F$ denotes the Frobenius norm. For $A = (a_{ij})^n_{i,j=1}$, $||A||^2_F = \mathrm{tr}(A^T A) = \sum^n_{i,j=1} a^2_{ij}$. This distance is computationally efficient but simply treats A, B as vectors in $\mathbb{R}^{n^2}$, disregarding all of their intrinsic structures (e.g. symmetry, positivity).
2. Square root distance [5] $d_{1/2}(A, B) = ||A^{1/2} - B^{1/2}||_F = (\mathrm{tr}[A + B - 2(A^{1/2}B^{1/2})])^{1/2}$.
3. Bures-Wasserstein distance $d_{\mathrm{BW}}(A, B) = (\mathrm{tr}[A + B - 2(B^{1/2}AB^{1/2})^{1/2}])^{1/2}$ (see e.g. [4, 9, 8, 3, 12]. This is precisely the $\mathcal{L}^2$-Wasserstein distance between two zero-mean Gaussian probability measures in $\mathbb{R}^n$ with covariance matrices A, B. It coincides with the square root distance if and only if A and B commute.

Consider now the set $\mathrm{Sym}^{++}(n)$ of $n \times n$ real, symmetric, positive definite (SPD) matrices. Elements of this set include, e.g., covariance matrices of Gaussian densities on $\mathbb{R}^n$. In general, by regularizing covariance matrices if necessary, we can treat them as elements of $\mathrm{Sym}^{++}(n)$. The set $\mathrm{Sym}^{++}(n)$ is rich in intrinsic geometrical structures and one common approach is to view it as a Riemannian manifold. Examples of Riemannian metrics that have been studied on $\mathrm{Sym}^{++}(n)$ include

1. Affine-invariant Riemannian metric (see e.g. [22, 2]), with the corresponding Riemannian distance $d_{\mathrm{aiE}}(A, B) = ||\log(B^{-1/2}AB^{-1/2})||_F$, where log denotes

the principal logarithm of A. Let $A = U\text{diag}(\lambda_1, \dots, \lambda_n)U^T$ be the spectral decomposition of A, then $\log(A) = U\text{diag}(\log(\lambda_1), \dots, \log(\lambda_n))U^T$. The affine-invariant Riemannian distance $d_{\text{aiE}}(A, B)$ corresponds to the Fisher-Rao distance between two zero-mean Gaussian densities with covariance matrices A, B in $\mathbb{R}^n$.

2. Log-Euclidean metric [1], with the corresponding Riemannian distance being the Log-Euclidean distance given by $d_{\text{logE}}(A, B) = ||\log(A) - \log(B)||_F$.
3. When restricted on $\text{Sym}^{++}(n)$, the Bures-Wasserstein distance is also the Riemannian distance corresponding to a Riemannian metric.

On $\text{Sym}^{++}(n)$, the Frobenius, square root, and Log-Euclidean distances are all special cases of the *power-Euclidean distances* [5], $d_{E,\alpha}(A, B) = \left\|\frac{A^\alpha - B^\alpha}{\alpha}\right\|, \alpha \in \mathbb{R}, \alpha \neq 0$, with $\lim_{\alpha \to 0} d_{E,\alpha}(A, B) = ||\log(A) - \log(B)||_F$. Here $A^\alpha = \exp(\alpha \log(A))$, where $\exp(A) = \sum_{k=0}^{\infty} \frac{A^k}{k!}$. Similarly, the Bures-Wasserstein and Log-Euclidean distances are special cases of the *α-Procrustes distances*, which are Riemannian distances corresponding to a family of Riemannian metrics on $\text{Sym}^{++}(n)$ [19, 17], $d^\alpha_{\text{proE}}(A, B) = \frac{(\text{tr}[A^{2\alpha} + B^{2\alpha} - 2(B^\alpha A^{2\alpha} B^\alpha)^{1/2}]^{1/2}}{|\alpha|}, \alpha \in \mathbb{R}, \alpha \neq 0$, with $\lim_{\alpha \to 0} d^\alpha_{\text{proE}}(A, B) = ||\log(A) - \log(B)||_F$. For a fixed $\alpha \neq 0$, the α-Procrustes distance coincides with the power-Euclidean distance if and only if A and B commute.

Infinite-dimensional generalizations. Consider now the setting of infinite-dimensional covariance operators. The finite-dimensional Frobenius distance generalizes readily to the infinite-dimensional Hilbert-Schmidt distance $||A - B||_{\text{HS}}$, where A, B are Hilbert-Schmidt operators. Similarly, the formulas for the square root and Bures-Wasserstein distances remain valid in the infinite-dimensional setting, where A, B are positive trace class operators on a Hilbert space.

The situation is substantially different with the Log-Euclidean and affine-invariant Riemannian distances (see also the discussion in [24]). This is due to the fact that a positive compact operator A on a Hilbert space, such as a covariance operator, possesses a sequence of eigenvalues approaching zero, and hence both A^{-1} and $\log(A)$ are unbounded. Thus the formulas for the Log-Euclidean and affine-invariant Riemannian distances cannot be carried over directly to the covariance operator setting. Instead, a proper infinite-dimensional generalization of the affine-invariant Riemannian and Log-Euclidean metrics on the set of SPD matrices have been proposed by using the concepts of *extended (unitized) Hilbert-Schmidt operators*, *positive definite (unitized) Hilbert-Schmidt operators*, and *extended Hilbert-Schmidt inner product and norm* [11]. We next discuss these concepts.

24.4 Riemannian Distances between Positive Definite Hilbert-Schmidt Operators

We first discuss the concept of *positive definite (unitized) Hilbert-Schmidt operators* on a Hilbert space [11]. Specifically, let $\mathcal{H}$ be an infinite-dimensional separable real Hilbert space with inner product $\langle\ ,\ \rangle$ and corresponding norm $||\ ||$. Let $\mathcal{L}(\mathcal{H})$ denote the set of bounded linear operators on $\mathcal{H}$. We recall the set of trace class operators on $\mathcal{H}$, $\text{Tr}(\mathcal{H}) = \{A \in \mathcal{L}(\mathcal{H}) : ||A||_{\text{tr}} = \sum_{k=1}^{\infty} \langle e_k, (A^*A)^{1/2} e_k \rangle < \infty\}$, where A^* is the adjoint of A, $\{e_k\}_{k=1}^{\infty}$ is any orthonormal basis in $\mathcal{H}$ and the definition of the trace norm $|| \, ||_{\text{tr}}$ is independent of the choice of such basis. For $A \in \text{Tr}(\mathcal{H})$, the trace of A is $\text{tr}(A) = \sum_{k=1}^{\infty}\langle e_k, Ae_k\rangle = \sum_{k=1}^{\infty} \lambda_k$, where $\{\lambda_k\}_{k=1}^{\infty}$ denote the eigenvalues of A. The set of Hilbert-Schmidt operators on $\mathcal{H}$ is defined to be $\text{HS}(\mathcal{H}) = \{A \in \mathcal{L}(\mathcal{H}) : ||A||_{\text{HS}}^2 = \text{tr}(A^*A) = \sum_{k=1}^{\infty} ||Ae_k||^2 < \infty\}$, the Hilbert-Schmidt norm $||A||_{\text{HS}}$ being independent of the choice of basis $\{e_k\}_{k=1}^{\infty}$. This set is itself a Hilbert space with the Hilbert-Schmidt inner product $\langle A, B\rangle_{\text{HS}} = \text{tr}(A^*B) = \sum_{k=1}^{\infty}\langle Ae_k, Be_k\rangle$. The set of *extended (or unitized) Hilbert-Schmidt operators* on $\mathcal{H}$ is defined in [11] to be

$$\text{HS}_X(\mathcal{H}) = \{A + \gamma I : A \in \text{HS}(\mathcal{H}), \gamma \in \mathbb{R}\}. \tag{24.3}$$

This is a Hilbert space under the *extended Hilbert-Schmidt inner product*

$$\langle A + \gamma I, B + \nu I\rangle_{\text{HS}_\text{X}} = \langle A, B\rangle_{\text{HS}} + \gamma\nu, \tag{24.4}$$

under which the scalar operators γI, $\gamma \in \mathbb{R}$, are orthogonal to the Hilbert-Schmidt operators. The corresponding *extended Hilbert-Schmidt norm* is given by

$$||A + \gamma I||_{\text{HS}_\text{X}}^2 = ||A||_{\text{HS}}^2 + \gamma^2, \tag{24.5}$$

under which $||I||_{\text{HS}_\text{X}} = 1$, in contrast to the Hilbert-Schmidt norm, where $||I||_{\text{HS}} = \infty$.

We recall that $A \in \mathcal{L}(\mathcal{H})$ is said to be *positive definite* [23] if $\exists M_A > 0$ such that $\langle x, Ax\rangle \geq M_A||x||^2 \; \forall x \in \mathcal{H}$. This condition is equivalent to requiring that A be both *strictly positive*, that is $\langle x, Ax\rangle > 0 \; \forall x \neq 0$, and *invertible*, with $A^{-1} \in \mathcal{L}(\mathcal{H})$. Let $\mathbb{P}(\mathcal{H})$ be the set of *self-adjoint positive definite* bounded operators on $\mathcal{H}$.

Positive definite (unitized) Hilbert-Schmidt operators. The set of *positive definite (unitized) Hilbert-Schmidt* operators on $\mathcal{H}$ is then defined to be

$$\begin{aligned}\mathscr{P}\mathscr{C}_2(\mathcal{H}) &= \mathbb{P}(\mathcal{H}) \cap \text{HS}_X(\mathcal{H}) \\ &= \{A + \gamma I \; : \; A \in \text{HS}(\mathcal{H}), A^* = A, \gamma \in \mathbb{R}, A + \gamma I > 0\}.\end{aligned} \tag{24.6}$$

This is a Hilbert manifold, being an open subset of the Hilbert space $\text{HS}_X(\mathcal{H})$. On $\mathscr{P}\mathscr{C}_2(\mathcal{H})$, both $\log(A + \gamma I)$ and $(A + \gamma I)^{-1}$ are well-defined and bounded.

Affine-invariant Riemannian distance. The generalization of the *affine-Riemannian metric* on $\text{Sym}^{++}(n)$ to the Hilbert manifold $\mathscr{P}\mathscr{C}_2(\mathcal{H})$ was defined in [11], with the corresponding Riemannian distance given by

$$d_{\text{aiHS}}[(A + \gamma I), (B + \nu I)] = ||\log[(B + \nu I)^{-1/2}(A + \gamma I)(B + \nu I)^{-1/2}]||_{\text{HS}_\text{X}}. \tag{24.7}$$

Log-Hilbert-Schmidt distance. Similarly, the generalization of the Log-Euclidean metric on $\text{Sym}^{++}(n)$ to $\mathscr{P}\mathscr{C}_2(\mathcal{H})$ was defined in [20], with the corresponding *Log-Hilbert-Schmidt distance* given by

$$d_{\text{logHS}}[(A + \gamma I), (B + \nu I)] = ||\log(A + \gamma I) - \log(B + \nu I)||_{\text{HS}_\text{X}}. \tag{24.8}$$

The definition of the extended Hilbert-Schmidt norm guarantees that both $d_{\mathrm{aiHS}}[(A+\gamma I),(B+\nu I)]$ and $d_{\mathrm{logHS}}[(A+\gamma I),(B+\nu I)]$ are always well-defined and finite for any pair $(A+\gamma I),(B+\nu I)\in\mathscr{PC}_2(\mathcal{H})$. In the setting of reproducing kernel Hilbert space (RKHS) covariance operators, both the affine-invariant Riemannian and Log-Hilbert-Schmidt distances admit closed form formulas in terms of the corresponding kernel Gram matrices [20], [14].

Distances between positive Hilbert-Schmidt operators. In the case $\gamma=\nu>0$ is fixed, both $d_{\mathrm{aiHS}}[(A+\gamma I),(B+\gamma I)]$ and $d_{\mathrm{logHS}}[(A+\gamma I),(B+\gamma I)]$ become distances on the set of self-adjoint, positive Hilbert-Schmidt operators on $\mathcal{H}$. In the following, let $\mathrm{Sym}(\mathcal{H})\subset\mathcal{L}(\mathcal{H})$ denote the set of self-adjoint, bounded operators and $\mathrm{Sym}^+(\mathcal{H})\subset\mathrm{Sym}(\mathcal{H})$ the set of self-adjoint, positive, bounded operators on $\mathcal{H}$.

Theorem 1 *Let $\gamma\in\mathbb{R},\gamma>0$ be fixed. The distances $d_{\mathrm{aiHS}}[(A+\gamma I),(B+\gamma I)]$, $d_{\mathrm{logHS}}[(A+\gamma I),(B+\gamma I)]$ are metrics on the set $\mathrm{Sym}^+(\mathcal{H})\cap\mathrm{HS}(\mathcal{H})$ of positive Hilbert-Schmidt operators on $\mathcal{H}$.*

Related and further generalizations. Similar to the extended Hilbert-Schmidt operators, we can define the *extended trace class operators* [15] to be $\mathrm{Tr}_X(\mathcal{H})=\{A+\gamma I: A\in\mathrm{Tr}(\mathcal{H}),\gamma\in\mathbb{R}\}$ along with the *extended Fredholm determinant* $\det_X(A+\gamma I)$ and subsequently the *extended Hilbert-Carleman determinant* [18]. With these concepts, we obtained the *infinite-dimensional Alpha Log-Det divergences* [15] and *Alpha–Beta Log-Det divergences* [16] between positive definite (unitized) trace class operators and subsequently on the entire Hilbert manifold $\mathscr{PC}_2(\mathcal{H})$ [18]. The Alpha-Beta Log-Det divergences form a highly general family of divergences on $\mathscr{PC}_2(\mathcal{H})$ and include the affine-invariant Riemannian distance as a special case.

The α**-Procrustes distances** can also be generalized to the infinite-dimensional setting of $\mathscr{PC}_2(\mathcal{H})$ and include both the Bures-Wasserstein and Log-Hilbert-Schmidt distances as special cases [19, 17],

$$\begin{aligned} & d^{\alpha}_{\mathrm{proHS}}[(A+\gamma I),(B+\gamma I)], \quad \alpha\in\mathbb{R},\alpha\neq 0 \qquad (24.9)\\ & =\frac{1}{|\alpha|}(\mathrm{tr}[(A+\gamma I)^{2\alpha}+(B+\gamma I)^{2\alpha}-2[(B+\gamma I)^{\alpha}(A+\gamma I)^{2\alpha}(B+\gamma I)^{\alpha}]^{1/2})])^{1/2},\\ & \lim_{\alpha\to 0} d^{\alpha}_{\mathrm{proHS}}[(A+\gamma I),(B+\gamma I)]=||\log(A+\gamma I)-\log(B+\gamma I)||_{\mathrm{HS_X}}. \qquad (24.10)\end{aligned}$$

In particular, for $A,B\in\mathrm{Sym}^+(\mathcal{H})\cap\mathrm{Tr}(\mathcal{H})$,

$$\lim_{\gamma\to 0} d^{1/2}_{\mathrm{proHS}}[(A+\gamma I),(B+\gamma I)]=2(\mathrm{tr}[A+B-2(B^{1/2}AB^{1/2})])^{1/2}. \qquad (24.11)$$

Finite-rank and finite-dimensional approximations. In practice, it is typically necessary to deal with *finite-rank* and/or *finite-dimensional approximations* of the above infinite-dimensional distances. In the cases of d_{aiHS} and d_{logHS}, these approximations are consequences of the following general convergence results.

Theorem 2 *Let $A,\{A_n\}_{n\in\mathbb{N}}\in\mathrm{Sym}(\mathcal{H})\cap\mathrm{HS}(\mathcal{H})$ be such that $\lim_{n\to\infty}||A_n-A||_{\mathrm{HS}}=0$. Assume that $(I+A)>0, I+A_n>0\ \forall n\in\mathbb{N}$. Then $\log(I+A_n)$, $\log(I+A)\in\mathrm{Sym}(\mathcal{H})\cap\mathrm{HS}(\mathcal{H})$ and*

$$\lim_{n\to\infty} ||\log(I+A_n) - \log(I+A)||_{\mathrm{HS}} = 0. \tag{24.12}$$

Theorem 3 *Let* $A, B, \{A_n\}_{n\in\mathbb{N}}, \{B_n\}_{n\in\mathbb{N}} \in \mathrm{Sym}(\mathcal{H})\cap\mathrm{HS}(\mathcal{H})$ *be such that* $\lim_{n\to\infty} ||A_n - A||_{\mathrm{HS}} = 0$, $\lim_{n\to\infty} ||B_n - B||_{\mathrm{HS}} = 0$. *Assume that* $(I+A) > 0, (I+B) > 0, I+A_n > 0, I+B_n > 0\ \forall n \in \mathbb{N}$. *Then* $(I+B_n)^{-1/2}(I+A_n)(I+B_n)^{-1/2} - I$, $(I+B)^{-1/2}(I+A)(I+B)^{-1/2} - I \in \mathrm{Sym}(\mathcal{H}) \cap \mathrm{HS}(\mathcal{H})$ *and*

$$\begin{aligned}\lim_{n\to\infty} ||&\log[(I+B_n)^{-1/2}(I+A_n)(I+B_n)^{-1/2}] \\ &- \log[(I+B)^{-1/2}(I+A)(I+B)^{-1/2}]||_{\mathrm{HS}} = 0.\end{aligned} \tag{24.13}$$

Finite-dimensional approximations via orthogonal projections. We now focus on the study of the finite-dimensional approximations of d_{aiHS} and d_{logHS} via orthogonal projections. Let $A \in \mathrm{HS}(\mathcal{H})$. Let $\{e_k\}_{k=1}^{\infty}$ be any orthonormal basis for $\mathcal{H}$. For any $f \in \mathcal{H}$, we have $f = \sum_{k=1}^{\infty}\langle f, e_k\rangle e_k$. Let $N \in \mathbb{N}$ be fixed and consider the finite-dimensional subspace $\mathcal{H}_N = \mathrm{span}\{e_k\}_{k=1}^{N}$. Consider next the projection operator $P_N = \sum_{k=1}^{N} e_k \otimes e_k : \mathcal{H} \to \mathcal{H}_N$. For any $f \in \mathcal{H}$, $P_N f = \sum_{k=1}^{N}\langle f, e_k\rangle e_k$ and for the operator $P_N A P_N : \mathcal{H} \to \mathcal{H}$,

$$P_N A P_N f = P_N \sum_{k=1}^{N}\langle f, e_k\rangle A e_k = \sum_{j=1}^{N}\left(\sum_{k=1}^{N}\langle f, e_k\rangle\langle Ae_k, e_j\rangle\right) e_j \in \mathcal{H}_N.$$

Thus $P_N A P_N$ is a finite rank operator, with rank at most N, and $\mathrm{range}(P_N A P_N) \subset \mathcal{H}_N$. In particular, $P_N A P_N|_{\mathcal{H}_N} : \mathcal{H}_N \to \mathcal{H}_N$ and for $f, g \in \mathcal{H}_N$, we have

$$\langle g, P_N A P_N f\rangle = \sum_{j,k=1}^{N}\langle f, e_k\rangle\langle g, e_j\rangle\langle Ae_k, e_j\rangle = \langle \mathbf{g}, \mathbf{A}_N\mathbf{f}\rangle_{\mathbb{R}^N}, \tag{24.14}$$

where $\mathbf{f} = (\langle f, e_k\rangle)_{k=1}^{N}$, $\mathbf{g} = (\langle g, e_k\rangle)_{k=1}^{N} \in \mathbb{R}^N$ and $\mathbf{A}_N$ is the $N \times N$ matrix with $(\mathbf{A}_N)_{ij} = \langle Ae_k, e_j\rangle$. Thus on $\mathcal{H}_N$ with basis $\{e_k\}_{k=1}^{N}$, $P_N A P_N|_{\mathcal{H}_N}$ is represented by the matrix $\mathbf{A}_N$. Furthermore, $A \in \mathrm{Sym}(\mathcal{H}) \Rightarrow P_N A P_N|_{\mathcal{H}_N} \in \mathrm{Sym}(\mathcal{H}_N) \Rightarrow \mathbf{A}_N \in \mathrm{Sym}(N)$ and $A \in \mathrm{Sym}^+(\mathcal{H}) \Rightarrow P_N A P_N|_{\mathcal{H}_N} \in \mathrm{Sym}^+(\mathcal{H}_N) \Rightarrow \mathbf{A}_N \in \mathrm{Sym}^+(N)$.

Theorem 4 ***(Finite-dimensional approximation of Log-Hilbert-Schmidt distance)*** *Assume that* $(A+I), (B+I) \in \mathscr{PC}_2(\mathcal{H})$. *Let* $A_N = P_N A P_N|_{\mathcal{H}_N}$ *and* $B = P_N B P_N|_{\mathcal{H}_N}$. *Then*

$$\lim_{N\to\infty} ||\log(A_N + I) - \log(B_N + I)||_{\mathrm{HS}} = ||\log(A+I) - \log(B+I)||_{\mathrm{HS}}. \tag{24.15}$$

Assume that $(A+\gamma I), (B+\gamma I) \in \mathscr{PC}_2(\mathcal{H})$, $\gamma \in \mathbb{R}, \gamma > 0$. *Then*

$$\lim_{N\to\infty} ||\log(A_N + \gamma I) - \log(B_N + \gamma I)||_{\mathrm{HS}} = ||\log(A+\gamma I) - \log(B+\gamma I)||_{\mathrm{HS}}. \tag{24.16}$$

Table 24.1 Classification errors on the test set

Distance	$(\sigma_1 = 1, \sigma_2 = 1.1)$	$(\sigma_1 = 1, \sigma_2 = 1.3)$	$(\sigma_1 = 1, \sigma_2 = 1.5)$
Hilbert-Schmidt	55%	22%	8%
square root	44%	6%	0%
Bures-Wasserstein	41%	6%	0%
Log-Hilbert-Schmidt	11%	0%	0%
Affine-invariant	16%	0%	0%

Theorem 5 ***(Finite-dimensional approximation of Affine-invariant Riemannian distance)*** *Assume that* $(A + I), (B + I) \in \mathscr{PC}_2(\mathcal{H})$. *Let* $A_N = P_N A P_N|_{\mathcal{H}_N}$ *and* $B = P_N B P_N|_{\mathcal{H}_N}$. *Then*

$$\lim_{N\to\infty} ||\log[(B_N + I)^{-1/2}(A_N + I)(B_N + I)^{-1/2}]||_{\text{HS}}$$
$$= ||\log[(B + I)^{-1/2}(A + I)(B + I)^{-1/2}]||_{\text{HS}}. \tag{24.17}$$

Assume that $(A + \gamma I), (B + \gamma I) \in \mathscr{PC}_2(\mathcal{H})$, $\gamma \in \mathbb{R}, \gamma > 0$. *Then*

$$\lim_{N\to\infty} ||\log[(B_N + \gamma I)^{-1/2}(A_N + \gamma I)(B_N + \gamma I)^{-1/2}]||_{\text{HS}}$$
$$= ||\log[(B + \gamma I)^{-1/2}(A + \gamma I)(B + \gamma I)^{-1/2}]||_{\text{HS}}. \tag{24.18}$$

24.5 Numerical Experiments on Gaussian Processes

We carry out the following binary classification of sample covariance operators corresponding to two zero-mean Gauss-Markov processes with covariance functions $\text{cov}_i(s,t) = \exp(-\sigma_i|s-t|)$, $\sigma_i > 0$, $i = 1, 2$, on the interval $I = [0, 1]$. For each process, we generated sample covariance operators, each using 500 sample paths on 201 regularly spaced points on $[0, 1]$. The training and testing sets contain 10 and 100 sample covariance operators, respectively, split equally between the two classes. For classification, we utilized the nearest neighbor approach. For the affine-invariant Riemannian and Log-Hilbert-Schmidt distances, we fixed $\gamma = 10^{-9}$. We reported the classification errors in three different scenarios in Table 24.1. For the setting $(\sigma_1 = 1, \sigma_2 = 1.5)$, the two Gaussian processes are easily distinguished and perfect classification is achieved for all except the Hilbert-Schmidt distance. For the case $(\sigma_1 = 1, \sigma_2 = 1.1)$, the two Gaussian processes are clearly much closer to each other and the distances performed differently, with the worst result by the Hilbert-Schmidt distance and the best result by the Log-Hilbert-Schmidt distance.

References

[1] Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Geometric means in a novel vector space structure on symmetric positive-definite matrices. SIAM J. on Matrix An. and App. **29**(1), 328–347 (2007)

[2] Bhatia, R.: Positive Definite Matrices. Princeton University Press (2007)

[3] Bhatia, R., Jain, T., Lim, Y.: On the Bures–Wasserstein distance between positive definite matrices. Expositiones Mathematicae (2018)

[4] Dowson, D., Landau, B.: The Fréchet distance between multivariate normal distributions. Journal of Multivariate Analysis **12**(3), 450 – 455 (1982)

[5] Dryden, I., Koloydenko, A., Zhou, D.: Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. Annals of Applied Statistics **3**, 1102–1123 (2009)

[6] Ferraty, F., Vieu, P.: Nonparametric functional data analysis: theory and practice. Springer (2006)

[7] Fremdt, S., Steinebach, J., Horváth, L., Kokoszka, P.: Testing the equality of covariance operators in functional samples. Scandinavian Journal of Statistics **40**(1), 138–152 (2013)

[8] Gelbrich, M.: On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces. Mathematische Nachrichten **147**(1), 185–203 (1990)

[9] Givens, C., Shortt, R.: A class of Wasserstein metrics for probability distributions. Michigan Math. J. **31**(2), 231–240 (1984)

[10] Horváth, L., Kokoszka, P.: Inference for Functional Data with Applications. Springer (2012)

[11] Larotonda, G.: Nonpositive curvature: A geometrical approach to Hilbert-Schmidt operators. Differential Geometry and its Applications **25**, 679–700 (2007)

[12] Malagò, L., Montrucchio, L., Pistone, G.: Wasserstein Riemannian geometry of Gaussian densities. Information Geometry **1**(2), 137–179 (2018)

[13] Masarotto, V., Panaretos, V., Zemel, Y.: Procrustes metrics on covariance operators and optimal transportation of Gaussian processes. Sankhya A pp. 1–42 (2018)

[14] Minh, H.: Affine-invariant Riemannian distance between infinite-dimensional covariance operators. In: International Conference on Geometric Science of Information (2015)

[15] Minh, H.: Infinite-dimensional Log-Determinant divergences between positive definite trace class operators. Linear Algebra and Its Applications **528**, 331–383 (2017)

[16] Minh, H.: Alpha-Beta Log-Determinant divergences between positive definite trace class operators. Information Geometry **2**(2), 101–176 (2019)

[17] Minh, H.: Alpha Procrustes metrics between positive definite operators: a unifying formulation for the Bures-Wasserstein and Log-Euclidean/Log-Hilbert-Schmidt metrics. arXiv:1908.09275 (2019)

[18] Minh, H.: Infinite-dimensional Log-Determinant divergences between positive definite Hilbert-Schmidt operators. Positivity (2019)

[19] Minh, H.: A unified formulation for the Bures-Wasserstein and Log-Euclidean/Log-Hilbert-Schmidt distances between positive definite operators. In: International Conference on Geometric Science of Information. Springer (2019)

[20] Minh, H.Q., Biagio, M.S., Murino, V.: Log-Hilbert-Schmidt metric between positive definite operators on Hilbert spaces. In: Advances in Neural Information Processing Systems 27 (NIPS 2014), pp. 388–396. Curran Associates, Inc. (2014)

[21] Panaretos, V., Kraus, D., Maddocks, J.: Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. Journal of the American Statistical Association **105**(490), 670–682 (2010)

[22] Pennec, X., Fillard, P., Ayache, N.: A Riemannian framework for tensor computing. International Journal of Computer Vision **66**(1), 41–66 (2006)

[23] Petryshyn, W.: Direct and iterative methods for the solution of linear operator equations in Hilbert spaces. Transactions of the American Mathematical Society **105**, 136–175 (1962)

[24] Pigoli, D., Aston, J., Dryden, I., Secchi, P.: Distances and inference for covariance operators. Biometrika **101**(2), 409–422 (2014)

[25] Ramsay, J., Silverman, B.: Functional data analysis. Springer (2005)

Chapter 25
Depth in Infinite-dimensional Spaces

Stanislav Nagy

Abstract Depth is a statistical tool that aims to introduce sensible data-dependent ordering of points in multivariate / function spaces. In theory, this should allow construction of statistical procedures based on ranks, orderings, or quantiles for multi-dimensional data. Some of the natural properties a depth should satisfy in finite-dimensional spaces however lose tractability and appeal as the dimension grows. We introduce the depth in finite-dimensional spaces, and outline particular difficulties one faces when attempting to generalize depths to the situation of functional, or other infinite-dimensional data.

25.1 Statistical Depth

Unlike for points in the real line, there is no natural ordering of d-dimensional vectors with $d > 1$. Therefore, the invaluable nonparametric statistical inference based on the notions of order statistics, ranks, and quantiles [6], breaks down when $\mathbb{R}^d$-valued random vectors are observed. An interesting solution to this problem was suggested in the 1970s by J. W. Tukey [15], who proposed to rank observations in $\mathbb{R}^d$ according to their centrality, as evaluated with respect to (w.r.t.) the given probability distribution P on $\mathbb{R}^d$. Denote by $\mathcal{P}(\mathcal{M})$ the set of all (Borel) probability measures on a measurable space $\mathcal{M}$, and suppose that we are given $x \in \mathbb{R}^d$ and $P \in \mathcal{P}\left(\mathbb{R}^d\right)$. The *halfspace* (or *Tukey*) *depth* of x w.r.t. P is defined as the smallest P-probability of a closed halfspace that contains x, that is

$$hD(x; P) = \inf_{u \in \mathbb{R}^d} \mathsf{P}\left(\langle X, u\rangle \leq \langle x, u\rangle\right). \tag{25.1}$$

Here, $X \sim P$ is a random vector with distribution P, and $(\Omega, \mathcal{A}, \mathsf{P})$ is the probability space on which all random quantities are defined.

Stanislav Nagy (✉)
Charles University, Faculty of Mathematics and Physics, Prague, Czech Rep.,
e-mail: nagy@karlin.mff.cuni.cz

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_25

The halfspace depth allows to order points of $\mathbb{R}^d$ in a distribution-specific manner. Given that the probability measure P is known, x is said to be “more centrally located" than y if $hD(x; P) > hD(y; P)$. This way, the point[1] $\widetilde{x} \in \mathbb{R}^d$ that satisfies $hD(\widetilde{x}; P) = \sup_{y \in \mathbb{R}^d} hD(y; P)$ serves as an analogue of the median for d-variate data, and is frequently called the *halfspace median* of P. Just as the median in $\mathbb{R}$, the halfspace median is known to have many beneficial properties; for instance, it is a quite robust location parameter, under mild conditions on P. On the other side of the spectrum, points whose halfspace depth $hD(\cdot; P)$ is small are the peripheral points that can be separated from the support of P by a hyperplane bounding a halfspace of small probability. Of course, in typical situations the true probability measure P is not known, and only a random sample $X_1, \dots, X_n$ from P is available. In that case we denote by $P_n \in \mathcal{P}(\mathbb{R}^d)$ the empirical measure of this random sample, and rank the points of $\mathbb{R}^d$ according to their sample depth $hD(\cdot; P_n)$. In the left panel of Fig. 25.1 we see several contours of the halfspace depth function for a random sample of bivariate data.

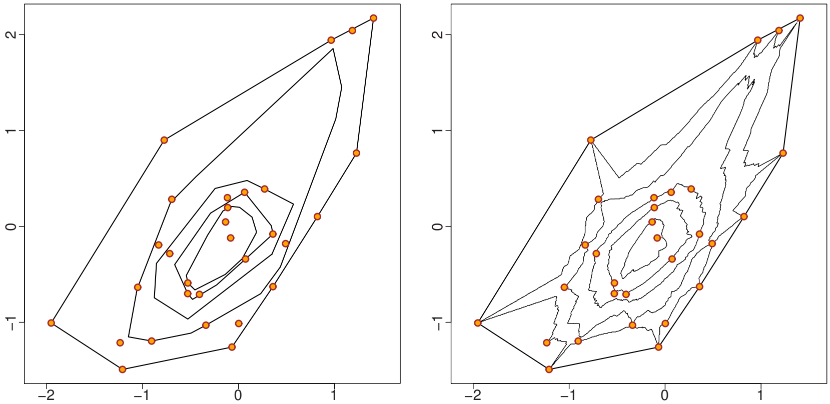


Fig. 25.1 A bivariate random sample of size 30 and several of its halfspace depth contours (left panel) and simplicial depth contours (right panel). For both concepts, the depth of a point outside the convex hull of the data is zero.

The halfspace depth is by far not the only depth available in $\mathbb{R}^d$. For instance, in 1988 the *simplical depth* was proposed [7] — for $x \in \mathbb{R}^d$ and $P \in \mathcal{P}(\mathbb{R}^d)$ we consider

$$sD(x; P) = \mathsf{P}(x \in \mathrm{co}(Z_1, \dots, Z_{d+1})) \tag{25.2}$$

where $Z_1, \dots, Z_{d+1}$ are independent realisations of $Z \sim P$, and $\mathrm{co}(\cdot)$ is the convex hull mapping. The simplical depth evaluates the probability that x is contained in a simplex whose vertices are randomly drawn from P.

[1] Or the barycentre of $\{x \in \mathbb{R}^d : hD(x; P) = \sup\{hD(y; P) : y \in \mathbb{R}^d\}\}$ if this set is not a singleton.

Desirable properties of finite-dimensional depths. Besides the renown halfspace and simplicial depth, hundreds of other depths are nowadays available in the literature. The first systematic treatment of the general concept of the depth comes with Y. Zuo and R. Serfling [16, 17], who argued that a reasonable *statistical depth function* $D\colon \mathbb{R}^d \times \mathcal{P}\left(\mathbb{R}^d\right) \to [0,1]\colon\ (x,P) \mapsto D(x;P)$ must satisfy (most of) the following conditions for all $P \in \mathcal{P}\left(\mathbb{R}^d\right)$:

(P1) *Affine invariance:* For any $A \in \mathbb{R}^{d\times d}$ non-singular and $b \in \mathbb{R}^d$, $D(x;P) = D\left(Ax+b; P_{AX+b}\right)$ for all $x \in \mathbb{R}^d$. Here, $X \sim P$, and $AX+b \sim P_{AX+b}$.
(P2) *Maximality at centre:* For P symmetric[2] around $\widetilde{x}$, $D\left(\widetilde{x};P\right) = \sup_{y\in\mathbb{R}^d} D\left(y;P\right)$.
(P3) *Decreasing along rays:* For $\widetilde{x}$ the depth median[3] of P, $x \in \mathbb{R}^d$ and $\lambda \in [0,1]$, $D(x;P) \leq D(\lambda x + (1-\lambda)\widetilde{x};P)$.
(P4) *Vanishing at infinity:* $\lim_{\|x\|\to\infty} D(x;P) = 0$.
(P5) *Semi-continuity in x:* Function $D(\cdot;P)$ is upper semi-continuous.
(P6) *Continuity in P:* As $P_\nu \to P$ weakly, $D(x;P_\nu) \to D(x;P)$ for all $x \in \mathbb{R}^d$.

Sometimes a condition stronger than **(P3)** is considered:

(P7) *Quasi-concavity in x:* All upper level sets of $D(\cdot;P)$ are convex.

The first four conditions are taken from [17], **(P5)**–**(P7)** are from [13]. It turns out that the halfspace depth satisfies all these conditions (with **(P6)** under a mild assumption on P). The simplicial depth violates **(P7)**, see also the right panel of Fig. 25.1 and [17].

All properties **(P1)**–**(P6)** are important. For random samples, however, especially **(P6)** is crucial, as it implies the consistency of the sample depth at x. In fact, for applications it is quite important that a uniform extension of **(P6)** holds true, at least for P_n empirical measures of random samples of size n from P:

(P8) *Uniform consistency:* $\sup_{x\in\mathbb{R}^d} |D(x;P_n) - D(x;P)| = 0$ almost surely.

Both hD and sD satisfy **(P8)**. In function spaces, an analogue of this property is extremely demanding, and separates the truly reasonable depths from other approaches.

Proposition 1 *For D that satisfies* ***(P1)*** *and* ***(P4)****, any $x \in \mathbb{R}^d$ outside the affine hull[4] of* $\operatorname{Supp}(P)$ *denoting the support of $P \in \mathcal{P}\left(\mathbb{R}^d\right)$, obtains zero depth, i.e. $D(x;P) = 0$.*

Proof If x lies outside the affine hull of $\operatorname{Supp}(P)$, there exists an affine transform A that maps $\operatorname{Supp}(P)$ into the hyperplane $H = \left\{y \in \mathbb{R}^d\colon y_d = 0\right\}$ and x into $A(x) = (0,\ldots,0,1)^\top \in \mathbb{R}^d$. A further linear map $y \mapsto (y_1,\ldots,y_{d-1},\lambda y_d)^\top$ with $\lambda > 0$ leaves H intact, yet translates $A(x)$ to $(0,\ldots,0,\lambda)^\top$. Altogether, by **(P1)** the depth of x w.r.t. P must equal the depth of $(0,\ldots,0,\lambda)^\top$ w.r.t. a fixed distribution supported in H, no matter which $\lambda > 0$ we chose. Condition **(P4)** allows to conclude. □

[2] A measure $P \in \mathcal{P}\left(\mathcal{M}\right)$ in a linear space $\mathcal{M}$ is (centrally) symmetric around $\widetilde{x} \in \mathcal{M}$ if for any $S \subset \mathcal{M}$ measurable we have $P(S - \widetilde{x}) = P(\widetilde{x} - S)$.

[3] The barycentre of the set of maximizers of $D(\cdot;P)$.

[4] Smallest translation of a vector subspace that contains the support of P.

The last observation will be useful in the study of functional depths. As we will see in Section 25.3, direct application of **(P1)–(P8)** fails in infinite-dimensional spaces.

25.2 Practice: Depth for Infinite-dimensional Data

Putting aside the discussion which of the properties **(P1)–(P8)** are reasonable to be expected from a depth in an infinite-dimensional space, let us first explore what kinds of depths for functional data have been considered in the literature. Our list is by no means exhaustive, but attempts to provide a fair outline of the state-of-the-art.

Suppose that B is an infinite-dimensional Banach space and $X \sim P \in \mathcal{P}(B)$. Typical examples of B we consider are the function spaces C of continuous functions from $[0,1]$ to $\mathbb{R}$, or the Hilbert space $\mathcal{L}^2$ of square-integrable curves on $[0,1]$.

Functional halfspace depth. A straightforward extension of a depth to B-valued data is a replacement of the inner product $\langle \cdot, u \rangle$ in (25.1) by a bounded linear functional $\phi \in B^*$ from the dual space B^* of B. This substitution yields the *functional halfspace depth* which to $x \in B$ and $X \sim P \in \mathcal{P}(B)$ assigns

$$hD(x;P) = \inf_{\phi \in B^*} \mathsf{P}\left(\phi(X) \leq \phi(x)\right), \tag{25.3}$$

considered for instance in [2]. Not much is known about the properties of this depth; it is usually discarded instantly because of the following observations.

Example 1 Consider $P \in \mathcal{P}(C)$ the distribution of the standard Wiener process symmetric around the constant zero function $0 \in C$. Immediately $hD(0;P) = 1/2$, and 0 is the only halfspace median of P. Nevertheless, for a random sample $X_1, \dots, X_n$ from P with empirical measure P_n it can be shown [4] that almost surely there exists $t_0 \in [0,1]$ such that $\min\{X_1(t_0), \dots, X_n(t_0)\} > 0$. Taking $\phi \in C^*$ the Dirac functional[5] at t_0, we get $hD(0;P_n) = 0$ almost surely for any $n = 1, 2, \dots$, and neither the analogue of **(P6)** nor **(P8)** can be true.

The situation is even worse — as shown in [2, Theorem 3] for a particular class of Gaussian processes, hD of almost all functions equals zero. Both these negative results are easily extended to other infinite-dimensional Banach spaces B. These problems with inconsistency and degeneracy are new in function spaces — we do not encounter them in $\mathbb{R}^d$. We will see later that they are inherently connected with the desired properties **(P1)** and **(P4)**.

Functional simplicial depth. An often considered extension of the simplicial depth to functional data is the *band depth* defined in the space C [8]. For $x \in C$ and $P \in \mathcal{P}(C)$ it takes the form[6]

[5] The Dirac functional at $t \in [0,1]$ is the evaluation mapping $\delta_t \colon C \to \mathbb{R} \colon x \mapsto x(t)$.

[6] For simplicity we consider only the depth from [8] for $J = 2$; extensions to $J > 2$ are straightforward, and share the same properties as this basic depth.

$$bD\,(x;P) = \mathsf{P}\,(\min\{Z_1(t), Z_2(t)\} \leq x(t) \leq \max\{Z_1(t), Z_2(t)\} \text{ for all } t \in [0,1]) \,.$$

Here Z_1, Z_2 are independent realisations of $Z \sim P \in \mathcal{P}\,(C)$. The band depth is a natural extension of sD, with the random simplex $\mathrm{co}\,(Z_1, \ldots, Z_{d+1})$ from (25.2) replaced by a random band of a pair of functions — the region in between the graphs of functions Z_1 and Z_2. Nevertheless, the band depth appears to suffer from the same problems as the functional halfspace depth. For functional data such as those from Example 1 it tends to degenerate, and fails to satisfy **(P8)** [4].

Integrated depths. A specific family of functional depths is covered by the umbrella term *integrated depths*. The original integrated depth from [3] was defined simply as an integral of (univariate) depths $hD(x(t); P_t)$ of functional values $x(t) \in \mathbb{R}$ of $x \in C$ at $t \in [0,1]$, w.r.t. the marginal distribution $P_t \in \mathcal{P}\,(\mathbb{R})$ of the functional value $X(t)$ of $X \sim P$; a comprehensive treatment of these depths can be found in [1]. In a general Banach space B, consider a (not necessarily probability) measure μ on the dual B^* of B. For any $\phi \in B^*$ we may project both $x \in B$ and $X \sim P \in \mathcal{P}\,(B)$ into $\mathbb{R}$, and subsequently a (univariate) depth of $\phi(x)$ can be evaluated w.r.t. the distribution of $\phi(X)$ denoted by $P_{\phi(X)} \in \mathcal{P}\,(\mathbb{R})$. The final ($\mu$-)*integrated depth* of x w.r.t. P is defined as the μ-integral of these depths of projected quantities

$$fD(x;P) = \int_{B^*} D\left(\phi(x); P_{\phi(X)}\right) \,\mathrm{d}\,\mu(\phi). \tag{25.4}$$

The original integrated depth from [3] is obtained by considering μ the uniform measure on the Dirac functionals $\{\delta_t : t \in [0,1]\}$. In general, fD evaluates the μ-weighted mean depth of a projection of x, w.r.t. the same projected quantity of $X \sim P$. Many depths given in the literature fall into the general framework of integrated depths. For instance, the popular *modified band depth* [8] is simply an integrated depth from [3] with D the simplicial depth (25.2) for one-dimensional data and μ as above.

Infimal depths. Suppose now that only a subset Φ of the dual space B^* is given. Instead of averaging the depths over projections from Φ, in [9] it was proposed to evaluate the minimum depth of a projection of x

$$iD\,(x;P) = \inf_{\phi \in \Phi} D\left(\phi(x); P_{\phi(X)}\right) . \tag{25.5}$$

This depth is called the (Φ-)*infimal depth* of x w.r.t. $X \sim P \in \mathcal{P}\,(B)$. It closely relates to the functional halfspace depth (25.3) — for D the halfspace depth (25.1) in $\mathbb{R}$, the infimum in (25.5) is taken over a subset Φ of all the functionals B^* defining halfspaces in B. Therefore, $iD(x;P) \geq hD(x;P)$ which should alleviate the problem with “too many projections" from Example 1. But, in fact already for the small set Φ of Dirac functionals $\{\delta_t : t \in [0,1]\}$, many of the undesirable properties of the functional halfspace depth remain to burden also the infimal depth. In particular, the

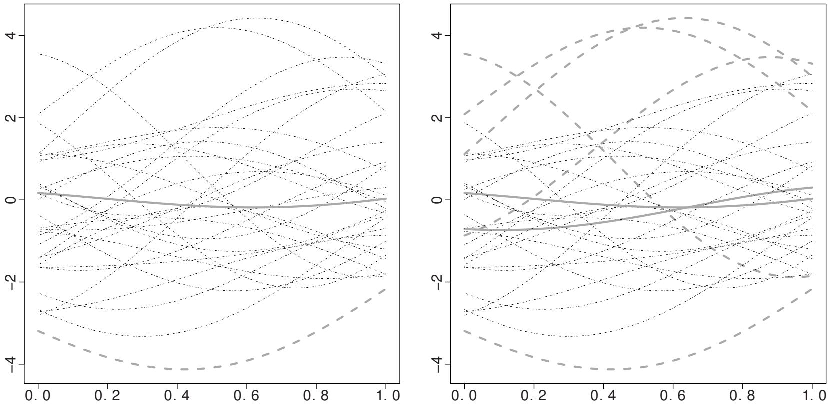


Fig. 25.2 A functional random sample of size 30 (all curves) along with the deepest sample function(s) (thick solid grey) and the least deep function(s) (dashed grey) for an integrated depth (left panel) and an infimal depth (right panel).

depth (25.5) still does not possess a universally consistent sample version [4], and tends to degenerate.[7]

Adaptive depths Integrated and infimal depths can be seen to come from a broader collection of depth functions. For $k \neq 0$ given, the *k-th moment integrated depth* [11] of $x \in \mathcal{L}^2$ w.r.t. $P \in \mathcal{P}\left(\mathcal{L}^2\right)$ is defined by

$$aD(x;P) = \left(\int_0^1 \left(D(x(t);P_t) + 1/2\right)^k \,\mathrm{d}t\right)^{1/k} - 1/2. \tag{25.6}$$

For $k = 1$ this depth equals the usual integrated depth (25.4) when applied as in [3]. Extensions of the k-th moment integrated depth towards other Banach spaces B, general subsets of projections $\Phi \subset B^*$, and different weighting measures μ as in (25.4) are straightforward. The advantage of considering the k-th power in (25.6) comes with its flexibility — as $k \to -\infty$, $aD(x;P)$ approaches the essential infimum of $D(x(t);P_t)$ over $t \in [0,1]$. This depth is essentially the infimal depth (25.5) (with Φ the Dirac functionals again). Therefore, the extended integrated depths encompass all the integrated, and all the infimal depths as special cases. Parameter k may be tuned to obtain versatile intermediate depths, similar in spirit to the extremal depth. This tuning allows to emphasize different properties of the underlying functions. Low k accentuates extremality, while higher k attaches more weight to the overall trend in the cross-sectional depth averaged over the domain. The choice of k may be data-driven; in many applications such as the classification of functional data this tuning appears to give quite promising results [11, Section 4]. Most importantly, the

[7] The concept of an infimal depth was rediscovered independently in [12]. There, the *extremal depth* similar to (25.5) was proposed with a refinement of a more elaborate tie-breaking procedure. The extremal depth, however, still fails to satisfy extensions of **(P6)** and **(P8)**.

adaptive depths (25.6) do not suffer from the shortcomings of the infimal depths; they do not degenerate, and satisfy **(P8)**, with no conditions needed for P [11, Section 3].

Finally, when the shape properties of functional data come into play, order-extensions of the adaptive depth (25.6) in the spirit of [10] can be introduced easily into (25.6). For details we refer to [11].

25.3 Theory: Desiderata for Depths in Function Spaces?

Having obtained several classes of depths that appear to work reasonably well in the practice of functional data analysis, let us turn to the more difficult question of which properties of a functional depth are desirable.

First, note that the finite-dimensional spaces $\mathbb{R}^d$ are all proper subspaces of $\mathcal{L}^2$, isometrically embedded for instance by the map

$$\mathbb{R}^d \to \mathcal{L}^2\colon \left(x = (x_1, \dots, x_d)^\mathsf{T}\right) \mapsto \left(t \mapsto \sqrt{d} \sum_{i=1}^{d} x_i \, \mathbb{I}\left[(i-1)/d \le t < i/d\right]\right).$$

If $\mathcal{L}^2$ is supposed to be considered with its linear structure, and if all the desired conditions **(P1)**–**(P8)** are expected to work also in $\mathcal{L}^2$, **(P1)** and **(P4)** imply that any curve $x \in \mathcal{L}^2$ outside the affine hull of $\mathrm{Supp}(P)$ gets zero depth. This follows from Proposition 1. For infinite-dimensional data, however, this entails degeneracy as encountered in Example 1. Indeed, for a random sample $X_1, \dots, X_n$ of Wiener processes the probability that any fixed function $x \in \mathcal{L}^2$ lies inside an affine hull of $X_1, \dots, X_n$ is zero. Likewise, $X_1, \dots, X_n$ are, with probability one, affinely independent, and using the proof of Proposition 1 again they must attain the same depth. Consequently, consistency **(P6)** and **(P8)** cannot be satisfied for non-trivial functional depths that satisfy extensions of **(P1)** and **(P4)**.

The main problem of **(P1)**–**(P8)** in functional settings appears to be the assumption of the linear structure, inappropriate for data of infinite dimensionality. While **(P8)** must be satisfied for any reasonable depth in any space,[8] the most crucial conditions **(P1)**, **(P3)** and **(P7)** must be revised. Currently it appears to be unclear how to replace these conditions, and more generally, what to expect from a depth in a Banach space.

Albeit analogues to **(P1)**–**(P8)** can be found in the literature [5], these have been cut to the bone, and present *minimal* requirements for a feasible depth.

So, what makes the depth for functional data so different from the depth in $\mathbb{R}^d$?

- *Every datum lives in its own subspace*. For infinite-dimensional random variables, it is typical that a random sample of n observations spans an n-dimensional affine space in B. The linear structure of B is thus inappropriate for inference.

[8] Compare with [14, Assertion 2.3].

- *Functions have shapes*. While the order of the coordinates in $\mathbb{R}^d$ is rather arbitrary,[9] the domain of functional data is naturally ordered. Functions may differ in shapes. This special trait is not addressed in most depth-based analysis.
- *Coordinate projections of functional data do not exactly match*. In functional depth it is assumed that the curves are aligned, meaning that no phase variation is allowed in P. This assumption is unrealistic, and it is unclear how to incorporate pre-alignment of the curves into the analysis based on the depth.

None of these challenges have been addressed adequately in the current literature on functional data depth. Despite the enormous advancement in the practice of functional depth and depth-based methods, our understanding of what the concept of depth means in truly infinite-dimensional spaces has hardly progressed over the past 20 years. This remains to be a major challenge in functional data analysis.

Acknowledgements This work was supported by the grant 19-16097Y of the Czech Science Foundation, and by the PRIMUS/17/SCI/3 project of Charles University.

References

[1] Cuevas, A., Fraiman, R.: On depth measures and dual statistics. A methodology for dealing with general data. J. Multivariate Anal. **100**(4), 753–766 (2009)

[2] Dutta, S., Ghosh, A.K., Chaudhuri, P.: Some intriguing properties of Tukey's half-space depth. Bernoulli **17**(4), 1420–1434 (2011)

[3] Fraiman, R., Muniz, G.: Trimmed means for functional data. Test **10**(2), 419–440 (2001)

[4] Gijbels, I., Nagy, S.: Consistency of non-integrated depths for functional data. J. Multivariate Anal. **140**, 259–282 (2015)

[5] Gijbels, I., Nagy, S.: On a general definition of depth for functional data. Statist. Sci. **32**(4), 630–639 (2017)

[6] Hájek, J., Šidák, Z., Sen, P.K.: Theory of rank tests. Probability and Mathematical Statistics, 2nd ed. Academic Press, Inc., San Diego (1999)

[7] Liu, R.Y.: On a notion of simplicial depth. Proc. Natl. Acad. Sci. U.S.A. **85**(6), 1732–1734 (1988)

[8] López-Pintado, S., Romo, J.: On the concept of depth for functional data. J. Amer. Statist. Assoc. **104**(486), 718–734 (2009)

[9] Mosler, K.: Depth statistics. In: C. Becker, R. Fried, and S. Kuhnt, editors, Robustness and complex data structures, pp. 17–34. Springer, Heidelberg (2013)

[10] Nagy, S., Gijbels, I., Hlubinka, D.: Depth-based recognition of shape outlying functions. J. Comput. Graph. Statist. **26**(4), 883–893 (2017)

[11] Nagy, S., Helander, S., Van Bever, G., Viitasaari, L., Ilmonen, P.: Adaptive integrated functional depths. Under review (2019)

[9] Inference in $\mathbb{R}^d$ is not affected by simultaneous permutations of all the coordinates of all the points.

[12] Narisetty, N.N., Nair, V.N.: Extremal depth for functional data and applications. J. Amer. Statist. Assoc. **111**(516), 1705–1714 (2016)

[13] Serfling, R.: Depth functions in nonparametric multivariate inference. In: Data depth: robust multivariate analysis, computational geometry and applications, volume 72 of DIMACS Ser. Discrete Math. Theoret. Comput. Sci., pp. 1–16. Amer. Math. Soc., Providence (2006)

[14] Serfling, R.: Depth functions on general data spaces, I. (2019) https://personal.utdallas.edu/~serfling/papers/I.Perspectives.pdf, cited 24 Jan 2020

[15] Tukey, J.W.: Mathematics and the picturing of data. In: Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), Vol. 2, pp. 523–531. Canad. Math. Congress, Montreal, (1975)

[16] Zuo, Y.: Contributions to the theory and applications of statistical depth functions. Ph.D. thesis, The University of Texas at Dallas (1998)

[17] Zuo, Y., Serfling, R.: General notions of statistical depth function. Ann. Statist. **28**(2), 461–482 (2000)

Chapter 26
Variable Selection in Semiparametric Bi-functional Models

Silvia Novo, Germán Aneiros and Philippe Vieu

Abstract A new sparse semiparametric functional model is proposed, which tries to incorporate the influence of two functional variables in a scalar response in a flexible way, but involving interpretable parameters. One of the functional variables is included trough a single-index structure and the other one linearly, but trough the high-dimensional vector formed by its discretized observations. For this model, a new algorithm for variable selection in the linear part is proposed. This procedure takes advantage of the functional origin of the scalar covariates with linear effect. Some asymptotic results will ensure the good performance of the method. Finally, Tecator's data will illustrate the great applicability of the presented methodology: good predictive power together with interpretability of the outputs.

26.1 Introduction

Functional variables are more and more common in practical situations and developing techniques with high level of flexibility and interpretability has became a target in current statistical researches. To be adapted to these practical requirements, models and procedures able to reduce dimensionality are of first necessity (see [9]) and both semiparametric and sparse ideas are of great interest for reaching this purpose.

In some situations, we have a scalar variable of interest, let say Y, and we want to know which points of the grid in which is observed a functional variable, namely $\zeta(t)$, are the most influential (points of impact) on this scalar variable. In other words, we

Silvia Novo (✉)
MODES research group, CITIC, Universidade da Coruña, A Coruña, Spain,
e-mail: s.novo@udc.es

Germán Aneiros
MODES research group, CITIC, ITMATI, Universidade da Coruña, A Coruña, Spain,
e-mail: german.aneiros@udc.es

Philippe Vieu
Institut de Mathématiques, Université Paul Sabatier, Toulouse, France,
e-mail: philippe.vieu@math.univ-toulouse.fr

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_26

want to select the relevant variables among the set of discretized observations of ζ. The problem is that standard variable selection methods, coming from an adaptation of the multivariate methodology, can provide inadequate results. On the one hand, these procedures are affected by the strong dependence between variables, which in this case is directly derived from its functional origin. On the other hand, the great quantity of observations makes difficult obtaining results in reasonable amount of time.

In [3], a new method is presented, the partitioning variable selection (PVS) procedure, for selecting impact points in the linear model

$$Y = a + \sum_{j=1}^{p_n} \beta_{0j}\zeta(t_j) + \varepsilon, \tag{26.1}$$

where ζ is a random curve defined on some interval $[a, b]$ and is observed in the points $a \leq t_1 < \cdots < t_{p_n} \leq b$ and ε denotes the random error. The main idea of the PVS method is creating a two-stage algorithm for selecting relevant variables, taking advantage of the fact that the covariates with linear effect come from a discretization of a curve. In this case, variables that are close in the discretization will contain very similar information on the response.

In [2], the PVS procedure has been extended to the semi-functional partial linear model (SFPLM), which is defined as

$$Y = \sum_{j=1}^{p_n} \beta_{0j}\zeta(t_j) + m(\mathcal{X}) + \varepsilon, \tag{26.2}$$

where $\mathcal{X}$ denotes a random variable valued on some separable Hilbert space, $\mathcal{H}$, and m is an unknown link function. However, practical requirements of controlling dimensionality and of associating interpretable parameters to both functional objects lead us to propose a new model.

Specifically, this paper focuses on a model based on a mixture of partial linear and single index ideas, the so-called semi-functional partial linear single index model (SFPLSIM), which is defined by the relationship:

$$Y = \sum_{j=1}^{p_n} \beta_{0j}\zeta(t_j) + m\left(\langle\theta_0, \mathcal{X}\rangle\right) + \varepsilon, \tag{26.3}$$

where θ_0 is an unknown functional direction in $\mathcal{H}$ and $\langle\cdot,\cdot\rangle$ denotes the inner product in this space.

In model (26.3) (as in models (26.1) and (26.2)), we assume that only a few scalar variables among the set $\{\zeta(t_1), \ldots, \zeta(t_{p_n})\}$ are going to form part of the model, so we are going to adapt the PVS methodology for selecting the relevant ones. From now on, let us denote as $S_n = \{j = 1, \ldots, p_n,\ \beta_{0j} \neq 0\}$ the set of subscripts corresponding to relevant variables and denote as $s_n = \sharp(S_n)$ its cardinal.

26.2 The PVS Procedure

For carrying out the two stages of the method, assume that we have a statistical sample of size n, $\{(\zeta_i, \mathcal{X}_i, Y_i), \quad i = 1, \dots, n\}$ i.i.d. as $(\zeta, \mathcal{X}, Y)$ and split this sample into two independent subsamples, asymptotically of the same size $n_1 \sim n_2 \sim n/2$, each one to be used in one stage of the method. Let us denote such subsamples as:

$$\mathcal{E}^1 = \{(\zeta_i, \mathcal{X}_i, Y_i),\ i = 1, \dots, n_1\},\ \ \mathcal{E}^2 = \{(\zeta_i, \mathcal{X}_i, Y_i),\ i = n_1 + 1, \dots, n_1 + n_2 = n\}.$$

Furthermore, assume without loss of generality that the number of covariates with linear effect, p_n, can be factorized in the following way: $p_n = q_n w_n$ with q_n and w_n integers.

26.2.1 First Stage

The first step of the PVS method is based on considering a reduced model, with only w_n covariates with linear effect (covering the entire discretization interval of ζ) and directly discard the other covariates with linear effect (since they contain very similar information about the response) before applying a standard procedure of variable selection. For that, only subsample $\mathcal{E}^1$ is used. Specifically:

1. Consider the set of variables $\mathcal{R}_n^1 = \{\zeta(t_k^1),\ k = 1, \dots, w_n\}$, where $t_k^1 = t_{[(2k-1)q_n/2]}$ and $[z]$ denotes the smallest integer not less than $z \in \mathbb{R}$.
2. The following model with only w_n covariates with linear effect is obtained:

$$Y_i = \sum_{k=1}^{w_n} \beta_{0k}^1 \zeta_i(t_k^1) + m^1\left(\langle \theta_0^1, \mathcal{X}_i \rangle\right) + \varepsilon_i^1. \tag{26.4}$$

3. Apply to model (26.4) the standard procedure presented in [8], based on Penalized Least Squares (PLS). An estimator $(\widehat{\boldsymbol{\beta}}_0^1, \widehat{\theta}_0^1)$ of the pair $(\boldsymbol{\beta}_0^1, \theta_0^1)$ is obtained, where $\boldsymbol{\beta}_0^1 = (\beta_{01}^1, \dots, \beta_{0w_n}^1)^\top$. Then, $\zeta(t_k^1)$ is selected in $\mathcal{R}_n^1$ if, and only if, $\widehat{\beta}_{0k}^1 \neq 0$.

26.2.2 Second Stage

In the second step of the PVS method, variables in the neighbourhood of the selected ones in the first stage are included. For that, only the subsample $\mathcal{E}^2$ is considered. Specifically:

1. A new set of variables is considered $\mathcal{R}_n^2 = \bigcup_{\left\{k, \widehat{\beta}_{0k}^1 \neq 0\right\}} \left\{\zeta(t_{(k-1)q_n+1}), \dots, \zeta(t_{kq_n})\right\}$. Denoting by $r_n = \sharp(\mathcal{R}_n^2)$, variables in $\mathcal{R}_n^2$ can be renamed as $\mathcal{R}_n^2 = \left\{\zeta(t_1^2), \dots, \zeta(t_{r_n}^2)\right\}$ and the following model can be considered:

$$Y_i = \sum_{k=1}^{r_n} \beta_{0k}^2 \zeta_i(t_k^2) + m^2\left(\langle \theta_0^2, \mathcal{X}_i \rangle\right) + \varepsilon_i^2. \tag{26.5}$$

2. Again, the PLS method presented in [8] is applied, but now to model (26.5), obtaining an estimator $(\widehat{\boldsymbol{\beta}}_0^{\mathbf{2}}, \widehat{\theta}_0^{\mathbf{2}})$ of the pair $(\boldsymbol{\beta}_0^{\mathbf{2}}, \theta_0^{\mathbf{2}})$, where $\boldsymbol{\beta}_0^{\mathbf{2}} = (\beta_{01}^{\mathbf{2}}, \dots, \beta_{0r_n}^{\mathbf{2}})^\top$. Then, $\zeta(t_k^{\mathbf{2}})$ is selected in $\mathcal{R}_n^{\mathbf{2}}$ if, and only if, $\widehat{\beta}_{0k}^{\mathbf{2}} \neq 0$.

26.2.3 Final Selection and Model Estimate

At the end of the PVS procedure, a variable $\zeta(t_j) \in \{\zeta(t_1), \dots, \zeta(t_{p_n})\}$ is selected if and only if belongs to $\mathcal{R}_n^{\mathbf{2}}$ and its estimated coefficient in the second stage, said $\widehat{\beta}_{0k_j}^{\mathbf{2}}$, is non-null. Therefore, the following estimated set of relevant variables is obtained, $\widehat{S}_n = \{j = 1, \dots, p_n,$ such that $t_j = t_{k_j}^{\mathbf{2}}$ with $\zeta(t_{k_j}^{\mathbf{2}}) \in \mathcal{R}_n^{\mathbf{2}}$ and $\widehat{\beta}_{0k_j}^{\mathbf{2}} \neq 0\}$.

In this case, a natural way of obtaining estimators for $\boldsymbol{\beta}_0$ and θ_0 in model (26.3) is using the estimations obtained in the second stage of the algorithm, we mean: $\widehat{\beta}_{0j} = \widehat{\beta}_{0k_j}^{\mathbf{2}}$ if $j \in \widehat{S}_n$, and $\widehat{\beta}_{0j} = 0$ otherwise. In the same way, $\widehat{\theta}_0 = \widehat{\theta}_0^{\mathbf{2}}$.

Denoting by $\widehat{\boldsymbol{\beta}}_0$ the vector of estimated linear coefficients, an estimator of the function $m_{\theta_0}(\cdot) \equiv m(\langle \theta_0, \chi \rangle)$ can be obtained by smoothing the residuals of the parametric fit:

$$\widehat{m}_{\widehat{\theta}_0}(\chi) \equiv \widehat{m}\left(\left\langle \widehat{\theta}_0, \chi \right\rangle\right) = \frac{\sum_{i=1}^n \left(Y_i - \boldsymbol{\zeta}_i^\top \widehat{\boldsymbol{\beta}}_0\right) K\left(d_{\widehat{\theta}_0}\left(\chi, \mathcal{X}_i\right)/h\right)}{\sum_{i=1}^n K\left(d_{\widehat{\theta}_0}\left(\chi, \mathcal{X}_i\right)/h\right)},$$

where we have denoted $\boldsymbol{\zeta}_i = \left(\zeta_i(t_1), \dots, \zeta_i(t_{p_n})\right)^\top$ and $h > 0$ is a bandwidth, K is a kernel function and, for any $\theta \in \mathcal{H}$, $d_\theta(\cdot, \cdot)$ is the semimetric defined as $d_\theta(\chi, \chi') = |\langle \theta, \chi - \chi' \rangle|$ for each $\chi, \chi' \in \mathcal{H}$.

26.3 Summary of Theoretical Results

In this work, the model selection consistency as well as the corresponding rates of convergence of the estimators derived from the PVS procedure ($\widehat{\beta}_0$, $\widehat{\theta}_0$ and $\widehat{m}_\theta$) are obtained. In particular, under suitable assumptions, we proved that:

- $\mathbb{P}(\widehat{S}_n = S_n) \longrightarrow 1$ as $n \longrightarrow \infty$.
- $\exists \gamma \geq 0$ such that $||\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0|| = O_p(n^{-1/2} s_n^\gamma)$,
- $\exists d : \mathbb{R} \longrightarrow (0, \infty)$ such that $||\widehat{\theta}_0 - \theta_0|| = O_p(n^{-1} d(h) s_n^{\gamma - 3/2})$.
- $\sup_{\theta \in \Theta_n} \sup_{\chi \in C} \left\{\left|\widehat{m}_\theta(\chi) - m_{\theta_0}(\chi)\right|\right\} = O_p\left(a_n\right) + O_p\left(n^{-1/2} s_n^{\gamma+1/2}\right)$, where a_n is the rate of convergence of the regression estimator in a pure functional model and Θ_n is a ball centred in θ_0 (for details, see [8]).

26.4 Real Data Application

In this section, a benchmark data set in the nonparametric functional context is modelled thought different functional regression models including the SFPLSIM proposed in this paper. The results obtained show the usefulness of both the SFPLSIM and the PVS estimation procedure.

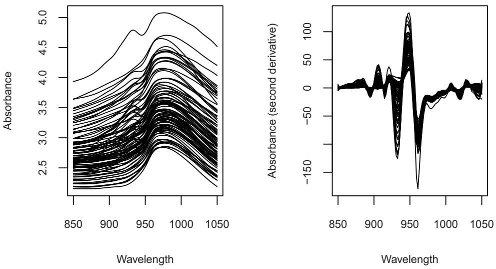


Fig. 26.1 Sample of 100 absorbance curves $\mathcal{X}$ (left panel) together with their second derivatives $\mathcal{X}^{(2)}$ (right panel)

Before beginning the next sections dedicated to present the data set, modelling and prediction, we indicate that, in the estimation of the two models that require variable selection (see models SFPLM and SFPLSIM), the tuning parameter of the penalty function (considered in the PLS procedure which is used in each stage of the PVS one) and h were selected by means of the BIC procedure, and the Epanechnikov kernel and the penalty function SCAD were used. In addition, in the SFPLSIM the order of the splines (in the spline basis for the estimation of θ_0) was $l = 3$, while the number of regularly interior knots, m_n, was fixed to 4 (since it is a moderate value that usually works well; big values of m_n have high computational cost). For details on the role of the splines, see [7].

26.4.1 Tecator's Data

The real data application will be focused on the well-known Tecator's data, which include the fat content and the near-infrared absorbance spectra of 215 finely chopped pieces of meat. For each piece of meat, the fat content, Y_i, is scalar, while the corresponding near-infrared absorbance spectra observations were collected on 100 equally spaced wavelengths (t_j, $j = 1, \ldots, 100$) in the range 850–1050 nm; so each subject can be considered as a continuous curve, $\mathcal{X}_i$. As usual when one deals with Tecator's data set, we will use the second derivatives of the absorbance curves, $\mathcal{X}_i^{(2)}$, as functional covariate instead of the original curve (see e.g. [6] for details). Figure 26.1 displays samples of both the absorbance curves and their second derivatives.

Our purpose is modelling the relation between the fat content and the absorbance spectra and then, use the model to predict the fat content. In order to compare the behaviour of each considered model and estimation procedure, we will split the original sample into two subsamples: a training sample, $\mathcal{T}_1 = \{(\mathcal{X}_i^{(2)}, Y_i)\}_{i=1}^{160}$, and a testing one, $\mathcal{T}_2 = \{(\mathcal{X}_i^{(2)}, Y_i)\}_{i=161}^{215}$. In this way, all the estimation task is made only by means of the training sample, while the testing sample is used to measure the quality of the predictions. To quantify the error in the prediction task, the mean square

Table 26.1 Values of the MSEP from some functional models

	Model	MSEP
FLM:	$Y = \alpha_0 + \int_{850}^{1050} \mathcal{X}^{(2)}(t)\alpha(t)dt + \varepsilon$	7.17
FNM:	$Y = r_1(\mathcal{X}^{(2)}) + \varepsilon$	4.06
FSIM:	$Y = r_2\left(\langle \theta_0, \mathcal{X}^{(2)} \rangle\right) + \varepsilon$	3.49

error of prediction (MSEP) will be used: MSEP $= \frac{1}{55}\sum_{i=161}^{215}\left(Y_i - \widehat{Y}_i\right)^2$, where $\widehat{Y}_i$ is the predicted value for Y_i obtained from each considered model and estimation procedure.

26.4.2 Modelling and Prediction Steps

In literature, several models have been used to describe the relation between the fat content and the absorbance spectra (see, for instance, [6] for a functional nonparametric model, and [5] for a multiple index functional model; see also [4] and [10] for functional partial linear and partial linear single-index models with exogenous covariates, respectively). [7] modelled this data set using the functional single-index model (FSIM) and compared the performance of the obtained predictions with that provided by the functional linear model (FLM) and the pure functional nonparametric model (FNM). Such three models as well as the corresponding MSEPs obtained from kernel estimation procedures are summarized in Table 26.1 (for details, see [7]). As can be observed, the relation between the fat content and the absorbance curves seems nonlinear (models with a nonparametric component, FNM and FSIM, offer much more accurate predictions than the linear model FLM).

However, more information can be taken from absorbance observations. For instance, the existence of points of impact in the spectrometric curves can be studied; that is, what values (if any) of the discretized curve, $(\mathcal{X}^{(2)}(t_1), \dots, \mathcal{X}^{(2)}(t_{100}))$, could improve the predictive results of FNM and FSIM.

Having in mind this idea, an extension of the FNM could be given by the SFPLM. In addition to the standard procedure of variable selection (PLS) proposed in [1] for the SFPLM, we will apply the PVS algorithm. Table 26.2 shows both the expression of the SFPLM considered and the MSEP results for the two variable selection methods (PLS and PVS). Note that it is obtained a clear improvement in the MSEP when one considers the SFPLM instead of the simpler model FNM (in fact, the SFPLM achieves even better performance than the FSIM). This improvement is even bigger when the PVS method is applied. Now, we put in practice the proposal in this paper: the PVS procedure applied to the SFPLSIM (note that this model can be seen as an extension of the FSIM). Both the expression of the SFPLSIM considered and the corresponding MSEPs are collected in Table 26.3. The better performance, from the point of view of the MSEP, of the SFPLSIM agaisnt the other four models considered in this section is evident (compare results in Tables 26.1-26.3). Furthermore, the PVS method overpasses the results of the PLS standard procedure, both in MSEP and simplicity of the model. The estimation of the functional direction, θ_0, using the

Table 26.2 Values of the MSEP from the SFPLM when the PLS or the PVS procedures are used (in parentheses, the number of covariates selected)

	Model	MSEP	
		PLS	PVS
SFPLM:	$Y=\sum_{j=1}^{100}\mathcal{X}^{(2)}(t_j)\beta_{0j}+m_1(\mathcal{X}^{(2)})+\varepsilon$	2.70 (8)	2.51 (10)

Table 26.3 MSEP when the proposed SFPLSIM and the PLS or the PVS procedures are used (in parentheses, the number of covariates selected)

	Model	MSEP	
		PLS	PVS
SFPLSIM:	$Y=\sum_{j=1}^{100}\mathcal{X}^{(2)}(t_j)\beta_{0j}+m\left(\left\langle\theta_0,\mathcal{X}^{(2)}\right\rangle\right)+\varepsilon$	0.96 (10)	0.88 (9)

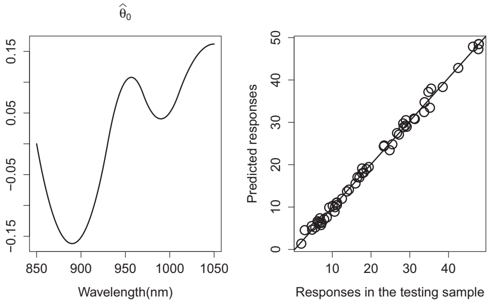


Fig. 26.2 Left panel: Estimate of the functional direction (θ_0) in the SFPLSIM. Right panel: Predicted values vs Observed values

PVS procedure is displayed in Figure 26.2 (left panel). It is worth being noted that the graphic of $\widehat{\theta}_0$ suggests that the two bumps around wavelengths 890 and 990, as well as a peak around wavelength 950, could be important indicators of the fat content. Finally, a graphic of the predicted values with the PVS procedure ($\widehat{Y}_i,\ i=161,\ldots,215$) versus the observed ones ($Y_i,\ i=161,\ldots,215$) can be seen in Figure 26.2 (right panel). The high predictive power of the SFPLSIM together with the PVS method is evident.

26.4.3 Summary

Our real data application evidences the advantages of using the SFPLSIM together with the PVS procedure in terms of accuracy of predictions. In addition, as in the case of FSIM, the SFPLSIM presents the advantage of the interpretation of the estimated direction of projection, $\widehat{\theta}_0$, which could also complement the information about how the (second derivative of the) spectrometric curves affect to the fat content.

Acknowledgements This work was supported in part by the Spanish Ministerio de Economía y Competitividad under Grant MTM2014-52876-R and Grant MTM2017-82724-R, in part by the Xunta de Galicia through Centro Singular de Investigación de Galicia accreditation under Grant ED431G/01 2016-2019 and through the Grupos de Referencia Competitiva under Grant ED431C2016-015, and in part by the European Union (European Regional Development Fund - ERDF). The first author also thanks the financial support from the Xunta de Galicia and the European Union (European Social Fund - ESF), the reference of which is ED481A-2018/191.

References

[1] Aneiros, G., Ferraty, F., Vieu, P.: Variable selection in partial linear regression with functional covariate. Stat. **49**(6), 1322–1347 (2015)

[2] Aneiros, G., Vieu, P.: Partial linear modelling with multi-functional covariates. Comput. Stat. **30**(3), 647–671 (2015)

[3] Aneiros, G., Vieu, P.: Variable selection in infinite-dimensional problems. Stat. Probab. Lett. **94**, 12–20 (2014)

[4] Aneiros, G., Vieu, P.: Semi-functional partial linear regression. Stat. Probab. Lett. **76**, 1102–1110 (2006)

[5] Chen, D., Hall, P., Müller, H.G.: Single and multiple index functional regression models with nonparametric link. Ann. Stat. **39**(3), 1720-1747 (2011)

[6] Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis: Theory and Practice. Springer series in Statistics, New York (2006)

[7] Novo, S., Aneiros, G., Vieu, P.: Automatic and location-adaptive estimation in functional single-index regression. J. Nonparametr. Stat. **31**(2), 364–392 (2019)

[8] Novo, S., Aneiros, G., Vieu, P.: Sparse semiparametric regression when predictors are mixture of functional and high-dimensional variables. Forthcoming (2020)

[9] Vieu, P.: On dimension reduction models for functional data. Statist. Probab. Lett. **136**, 134–138 (2018)

[10] Wang, G., Feng, X.N., Chen, M.: Functional partial linear single-index model. Scand. J. Stat. **43**, 261–274 (2016)

Chapter 27

Local Inference for Functional Data Controlling the Functional False Discovery Rate

Niels Lundtorp Olsen, Alessia Pini and Simone Vantini

Abstract A topic which is becoming more and more popular in Functional Data Analysis is local inference, i.e., the continuous statistical testing of a null hypothesis along a domain of interest. The principal issue in this topic is the infinite amount of tested hypotheses, which can be seen as an extreme case of multiple comparisons problem. A number of quantities have been introduced in the literature of multivariate analysis in relation to the multiple comparisons problem. Arguably the most popular one is the False Discovery Rate (FDR), that measures the expected proportion of false discoveries (rejected null hypotheses) among all discoveries. We define FDR in the setting of functional data defined on a compact set of $\mathbb{R}^d$. A continuous version of the Benjamini-Hochberg procedure is introduced, along with a definition of adjusted p-value function. Some general conditions are stated, under which the functional Benjamini-Hochberg (fBH) procedure provides control of FDR. We show how the procedure can be plugged-in with every parametric or nonparametric pointwise test, given that such test is exact. Finally, the proposed method - together with a nonparametric test - is applied to the analysis of the benchmark dataset of Canadian temperatures.

27.1 Introduction

In functional data analysis (FDA), the object of statistical methods are functions, which are typically modeled as random elements of a Hilbert space [6]. In this framework inference is particularly challenging, since it deals with elements of infinite dimensional spaces.

Niels Lundtorp Olsen
University of Copenhagen, Denmark, e-mail: niels.olsen@math.ku.dk

Alessia Pini (✉)
Università Cattolica del Sacro Cuore, Italy, e-mail: alessia.pini@unicatt.it

Simone Vantini
Politecnico di Milano, Italy, e-mail: simone.vantini@polimi.it

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_27

Literature in this field first focused on so-called ‘global’ methods [4, 5]: testing procedures where functions are treated as the atoms of the statistical analysis. In such a case, inference is performed by means of a unique test, resulting in a global p-value. For instance, if we are interested in comparing two groups of curves, the rejection of the null hypothesis of equality in mean or distribution between the two groups means that they differ significantly in at least one portion of the domain. However, no information is given on the particular portion responsible for such a rejection.

Lately, inferential methods for functional data have started focusing instead on ‘local’ techniques. In this case, inference is performed locally on the domain, and the identification of the areas of the domain responsible for the rejection of the null hypothesis is provided. Local inferential techniques are either based on simultaneous confidence bands, which are provided with a fixed coverage probability [3, 7], or on the definition of a p-value function, that provides a p-value at each point of the domain, guaranteeing a control of a quantity related with the error rate on the whole domain. Focusing on this second line of research, depending on the quantity that is controlled, different methods can be defined. Many papers deal with the extension of the control of the family-wise error rate (FWER) - a well known quantity defined for multivariate data - to the case of functional data. For instance, [12] propose a procedure controlling in a strong way the FWER between the elements of a partition, while [10] propose to control the FWER over intervals.

In this work we focus instead on the false discovery rate (FDR), that was first introduced in the seminal paper by Benjamini and Hochberg [1]. In particular, we describe an extension of the FDR to functional data, and the functional Benjamini-Hochberf (fBH) procedure: a procedure able to control the functional FDR. All details about the FDR-controlling procedure, as well as the proofs of all results described here are reported in [8]. Finally, to show the practical usefulness of our procedure, we show the application to a well know dataset of Canadian temperatures. Our procedure is applied to test differences in temperature distribution between different Canadian regions.

An complete description of the theoretical properties of the fBH procedure, its extension to the case of data defined on manifolds, an extensive simulation study comparing it with other state-of-the-art methods, and an application of fBH to a data set of daily temperatures on the Earth are reported in [8].

27.2 False Discovery Rate for Functional Data

FDR - first defined by Benjamini and Hochberg [1] - is a well-known target quantity in multiple testing. In the framework of testing multiple hypotheses, the FDR is defined as the expected proportion of correctly rejected null hypotheses among all rejected hypotheses. In [1], the Benjamini-Hochberg (BH) procedure is also provided, that is a procedure controlling the FDR under the condition of positive regression dependence on the subset of true null hypotheses (PDRS, see [1] and [2] for details).

In the case of functional data, local null and alternative hypotheses are defined for each point of the domain, so they are an infinite uncountable quantity. Hence, it does not make sense to talk about number of correctly or falsely rejected null hypotheses. In this section we show how it is possible to extend the notion of FDR to functional data, and we present a procedure controlling this quantity.

27.2.1 Definition of the Functional FDR

Consider a data set of continuous functions defined on the common domain $D \subset \mathbb{R}^d$, with $d \geq 1$. Assume that D is compact and measurable. For all $t \in D$, define as H_0^t and H_1^t a null and an alternative hypothesis, respectively, and let $p(t)$ denote the p-value of a test of H_0^t against H_1^t. The collection of $p(t)$ for all $t \in D$ is referred-to as the unadjusted p-value function [10].

Let U denote the portion of the domain where the null hypothesis is true: $U = \{t \in D : H_0^t \text{ is true}\}$. In the following, we assume that the test of H_0^t against H_1^t is exact, that is:

$$\forall \alpha \in (0, 1), \forall t \in U, \quad \mathbb{P}[p(t) \leq \alpha] = \alpha.$$

In order to be able to define the functional FDR, we first define the following quantities, related to the portions of the domain where the local null hypothesis is true/false, and where it is rejected.

Definition 1 Given U and an instance of $p(t)$, define the following subsets of the domain:

- $V = \{t : H_0^t \text{ is true and } H_0^t \text{ is rejected}\}$
- $S = \{t : H_0^t \text{ is false and } H_t^0 \text{ is rejected}\}$

The random set V corresponds to committing type I errors, and in a given research situation, it is desirable that V is as small as possible and S is as large as possible. The union of V and S gives the portion of the domain where the null hypothesis is rejected. In the FDR context, we are interested in controlling the expected proportion between the measure of the region where the null hypothesis is wrongly rejected V, and the measure of the region where the null hypothesis is rejected, that is $V \cup S$. So, we define the functional FDR (fFDR) as follows.

Definition 2 The *functional false discovery rate* (fFDR) is defined as

$$\text{fFDR} = \mathbb{E}[Q] = \mathbb{E}\left[\frac{\mu(V)}{\mu(V \cup S)} 1_{\mu(V \cup S)>0}\right]$$

where $Q = \frac{\mu(V)}{\mu(V \cup S)} 1_{\mu(V \cup S)>0}$ is the *proportion of false discoveries*, and μ is the Lebesgue measure.

Definition 2 is based on the Lebesgue measure of the regions V and S, but it can be extended to the case of any bounded measure of D that is absolutely continuous with respect to the Lebesgue measure. This extension can be used to deal with functional data defined over manifolds. See [8] for further details.

27.2.2 Control of the fFDR

We now focus on the control of the fFDR. As shown in [8] - and analogously to the multivariate case - we can define the functional Benjamini-Hochberg (fBH) procedure as the extension of the Benjamini-Hochberg procedure for functional data.

The fBH procedure can be defined by replacing the counts and sums in the original BH procedure with the Lebesgue measure μ.

Definition 3 (Functional Benjamini-Hochberg procedure - adjusted threshold)

Let $\alpha \in (0,1)$ be a desired significance level for the tests. The functional Benjamini-Hochberg (fBH) procedure is: *Reject hypotheses H_0^t that satisfy*

$$p(t) \leq \alpha^* \quad \text{where} \quad \alpha^* = \arg\max_r \left\{ \frac{\mu(\{s : p(s) \leq r\})}{\mu(D)} \geq \alpha^{-1} r \right\}$$

We will refer to α^* as the *adjusted threshold* of the procedure, and the function $a(r) = \mu(\{s : p(s) \leq r\})$ as the *cumulated p-value function*.

The fBH procedure of Definition 3 is based on adjusting the threshold α to apply to the unadjusted p-value function to select the portions of the domain where the null hypothesis is rejected. In the application of FDR-controlling procedures, it is often of interest to have the possibility of keeping the threshold fixed, and adjusting the p-value instead. As discussed by [8], the fBH procedure can be equivalently defined through the notion of fFDR-adjusted p-value function.

Definition 4 (Functional Benjamini-Hochberg procedure - adjusted p-value)

The fFDR-adjusted p-value function $\tilde{p}(t)$ is defined as:

$$\tilde{p}(t) = \min_{s \geq p(t)} \left\{ 1, \frac{\mu(D)s}{\mu(r : p(r) \leq s)} \right\}, \qquad t \in \mathrm{d}$$

It can be shown theoretically that the two definitions coincide. Furthermore, a finite-dimensional approximation of the procedure that is based on a grid-approximation of data can be given by applying multivariate BH procedure on the pointwise evaluations of functional data. Such approximation converges to the continuous version of the procedure when the sampling density goes to infinity. Finally, it can be shown that the fBH procedure controls the fFDR by $\alpha\mu(U)/\mu(D) \leq \alpha$ under some regularity conditions (see [8] for all details) and under PDRS assumption on data.

Note that what is required to apply the fBH procedure is to perform an exact pointwise test over the domain of interest, in order to obtain the p-value function $p(t)$. This means that it is possible to plug-in the procedure with every suitable test. In particular, if a nonparametric test is used, distributional assumptions on the functional data are not required, making the fBH procedure very flexible.

27.3 Analysis of Canadian Daily Temperatures

The aim of this section is to illustrate the potential of the fBH approach. We choose to apply the procedure to a well-known benchmark functional data set, i.e. the Canadian daily temperatures data set [11]. In addition, we compare the fBH procedure with the interval-wise testing (IWT) procedure, proposed by Pini and Vantini [10]. The data set contains the average daily temperatures (over 30 years) recorded by 35 weather stations in Canada. Coherently with previous analyses of the Canadian daily temperature data [11], functional data have been obtained by a standard Fourier smoothing on 65 harmonics.

The weather stations are divided into four climate zones: Atlantic, Pacific, Continental, and Arctic. We test the equality of the mean temperatures of the four climatic zones in a pairwise perspective with nonparametric permutation tests, as done in [10]. Denote as $y_{j1}(t), y_{j2}(t), \dots, y_{jn_j}(t)$ the functional data of the n_j stations of a climatic zone $j \in \{1, \dots, 4\}$. Assume that $\forall j : y_{ji}(t) = \mu_j(t) + \varepsilon_{ij}(t)$, and that $\varepsilon_{ij}(t)$ are independent and identically distributed random functions. We want to test for each couple $(j, j') \in \{1, \dots, 4\}^2$, $j \neq j'$ the following hypotheses:

$$H_0^t : \mu_j(t) = \mu_{j'}(t) \text{ against } H_1^t : \mu_j(t) \neq \mu_{j'}(t).$$

To perform the test of comparisons between climatic zones we employ nonparametric permutation tests [9]. Figure 27.1 reports the results of the analysis.

The hypothesis of equality on distribution between two climatic zones is tested by means of nonparametric permutation tests, based on a squared mean-difference test statistic. This leads to the computation of the unadjusted p-value function (solid black line on the left panels of Figure 27.1), as described in [10]. The fBH procedure defined in 4 is then applied to compute the fFDR adjusted p-value function (gray dashed line). Finally, the IWT is also performed on the same data, leading to the IWT-adjusted p-value function (gray dotted line).

It can be seen from Figure 27.1 that the fBH procedure is less conservative than IWT in the majority of comparisons (fBH-adjusted p-value is often lower than IWT-adjusted one). This is consistent with the different theoretical properties of the two methods: the fFDR control implies that the average proportion of false discoveries among the discoveries is below α. This type of control is often less conservative than the FWER control, which focuses on the probability of committing at least one false discovery on the whole domain. More in general, the results the two procedure provide very similar interpretation of data differences: both Atlantic and Pacific zones significantly differ from the Arctic zone over the entire year. Temperatures of these two zones also significantly differ from the Continental ones during winter, while Continental and Arctic zones are significantly different during the whole year but for winter months.

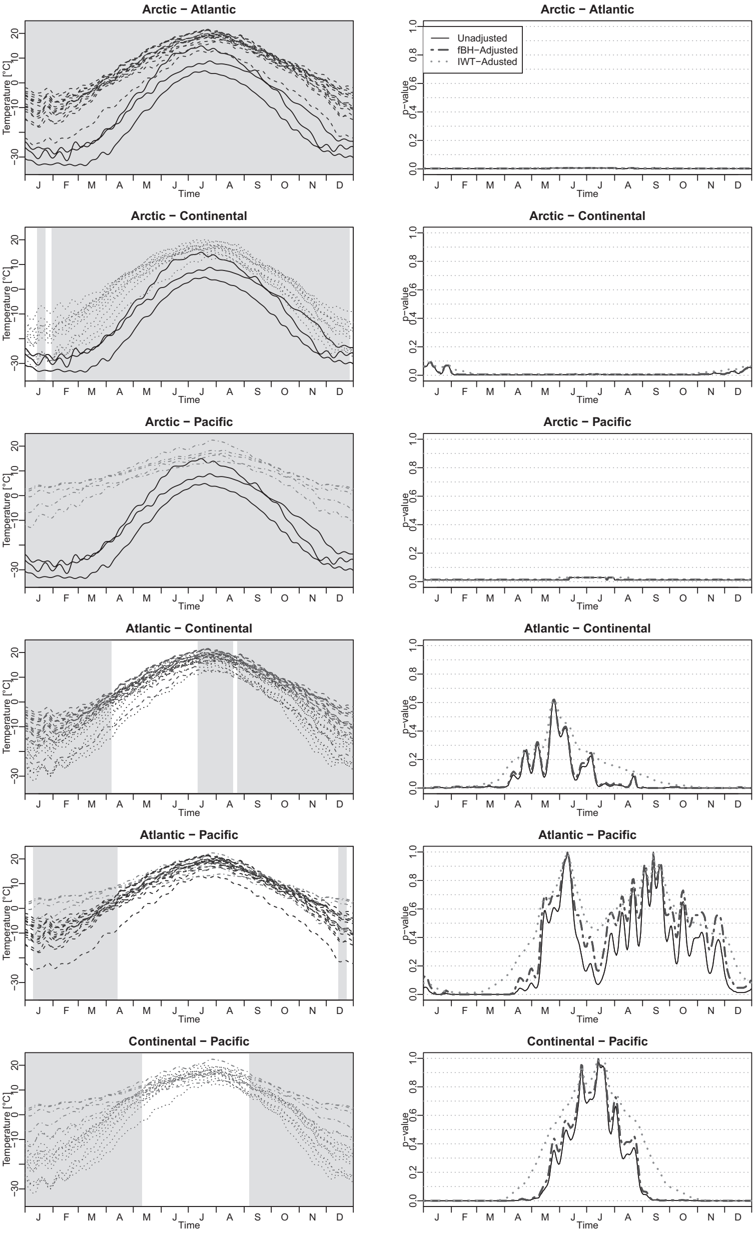


Fig. 27.1 Left: unadjusted, fFDR adjusted, and IWT-adjusted p-value functions associated to the pairwise differences between daily temperatures of four Canadian zones. Right: functional data and periods of the year with significant differences between each pair of climatic zones controlling the fFDR error rate at 5% (grey areas).

References

[1] Benjamini, Y., Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B **57**(1), 289–300 (1995)

[2] Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. Ann. Statist. **29**(4), 1165–1188 (2001)

[3] Choi, H., Reimherr, M.: A geometric approach to confidence regions and bands for functional parameters. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **80**(1), 239–260 (2018)

[4] Hall, P., Tajvidi, N.: Permutation tests for equality of distributions in high-dimensional settings. Biometrika **89**(2), 359–374 (2002)

[5] Horváth, L., Kokoszka., P.: Inference for functional data with applications. Springer (2012)

[6] Hsing, T., Eubank, R.: Theoretical foundations of functional data analysis, with an introduction to linear operators. John Wiley & Sons (2015)

[7] Liebl, D., Reimherr, M.: Fast and Fair Simultaneous Confidence Bands for Functional Parameters. arXiv:1910.00131 (2019)

[8] Lundtorp Olsen, N., Pini, A., Vantini, S.: False discovery rate for functional data. arXiv:1908.05272 (2019)

[9] Pesarin, F., Salmaso, L.: Permutation tests for complex data: theory, applications and software. John Wiley & Sons Inc (2010)

[10] Pini, A., Vantini, S.: Interval-wise testing for functional data. Journal of Nonparametric Statistics **29**(2): 407–424 (2017)

[11] Ramsay, J.O., Silverman., B.W.: Functional data analysis. Springer, New York (2005)

[12] Vsevolozhskaya, O., Greenwood, M., Holodov, D.: Pairwise comparison of treatment levels in functional analysis of variance with application to erythrocyte hemolysis. Ann. Appl. Stat. **8**(2), 905–925 (2014)

Chapter 28
Optimum Scale Selection for 3D Point Cloud Classification through Distance Correlation

Manuel Oviedo de la Fuente, Carlos Cabo, Celestino Ordóñez and Javier Roca-Pardiñas

Abstract Multiple scale machine learning algorithms using handcrafted features are among the most efficient methods for 3D point cloud supervised classification and segmentation. Despite their proven good performance, there are still some aspects that are not fully solved, determining optimum scales being one of them. In this work, we analyze the usefulness of functional distance correlation to address this problem. Specifically, we propose to adjust functions to the distance correlation between each of the features, at different scales, and the labels of the points, and select as optimum scales those corresponding to the global maximum of said functions. The method, which to the best of our knowledge has been proposed in this context for the first time, was applied to a benchmark dataset and the results analyzed and compared with those obtained using other methods for scale selection.

28.1 Introduction

In recent decades there has been an explosion of sensors and techniques to obtain spatial data representing real objects by means of 3D point clouds. Laser scanners, either static, mobile, portable or airborne, as well as cameras and computer vision algorithms, especially the Structure-from-Motion (SfM) algorithm [13], are currently the main sources of this kind of data. From the beginning, it was quite evident that there was a need to develop algorithms for the automatic extraction of useful information from the point clouds; a need that increased with the progressive capacity

Manuel Oviedo de la Fuente (✉)
Universidade de Santiago de Compostela, Spain, e-mail: manuel.oviedo@usc.es

Carlos Cabo
Universidad de Oviedo, Spain, e-mail: cabo.gmail@uniovi.es

Celestino Ordóñez
Universidad de Oviedo, Spain, e-mail: ordonezcelestino@uniovi.es

Javier Roca-Pardiñas
Universidade de Vigo, Spain, e-mail: roca@uvigo.es

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_28

of the sensors to measure larger point clouds each time. Among these algorithms, those based on the application of machine learning techniques have proven to be very efficient [10, 4], and, accordingly, their use is very extended nowadays. Moreover, their efficiency increases when the features (covariates) of the model are not extracted at a single scale but at several scales. In practice, this means that feature extraction is carried out considering different sizes of the neighborhood (scale) around each point (or voxel, when the point clouds is simplified by means of voxelization). In this way, the extracted features for a point (voxel) at different scales capture different characteristics of the objects around that point, and this helps in the classification procedure. Unfortunately, the selection of the scales is often carried out heuristically, taking into account the density of the point cloud, the kind of objects to be classified, the noise of the data, etc. On other ocassions, the procedure simply consists in selecting a number of scales at regular intervals. These procedures are quite objective, and have some drawbacks [7]. For that reason, it is important to carry out research into more objective scale selection methods, as an adequate selection has a positive influence on the results of the classification. Previous works have addressed this problem from different perspectives. One of them is to find the scale for which the labelling of the current point is the most similar to the labellings of its neighbors at the same scale [4]. Another approach [15] estimates the optimum scales taking into account the local structure of the covariance matrix and the Shanon entropy [11]. In this work, we propose a different approach that assumes that the optimum scales should correspond to the local maximum of the functions obtained calculating the distance correlation between each of the features at a number of scales (i.e. 100) and the values of the labels.

28.2 Methodology

28.2.1 Feature Extraction

A key aspect of machine learning applied to point cloud segmentation and classification is to define and determine the features (input variables) to be introduced in the mathematical models. The multi-scale strategy is based on the fact that a region around a point can look like a 1D, 2D or 3D object depending on the size of the region [2]. The input variables (features) included in the supervised classification algorithms are algebraic expressions involving the eigenvalues of the eigendecomposition of the local covariance matrix $\mathbf{\Sigma}$: $\Sigma = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{p}_i - \bar{\mathbf{p}})^T(\mathbf{p}_i - \bar{\mathbf{p}}) = \mathrm{V}\Lambda\mathrm{V}^{\mathrm{T}}$, where $\mathbf{p}_i = (X_i, Y_i, Z_i)$ is a point of the point cloud, $\lambda_1 > \lambda_2 > \lambda_3$ are the eigenvalues, V a matrix whose columns are the corresponding eigenvectors, and N the number of points inside a sphere of center $\mathbf{p}_i$ and radius R. That is, the eigenvalues, and consequently the features extracted from the point cloud, depend on the values of the scale (radius of the sphere).

The relationship between the values of the eigenvalues $\lambda_1, \lambda_2, \lambda_3$ at a point is related to the local geometry at that point [5]: a linear 1D structure when $\lambda_1 \geq \lambda_2, \lambda_3$; a planar 2D structure when $\lambda_1, \lambda_2 \geq \lambda_3$ and a volumetric 3D structure when

$\lambda_1 \approx \lambda_2 \approx \lambda_3$. Specifically, the features extracted for each point at each scale are: *Linearity* $L = (\lambda_1 - \lambda_2)/\lambda_1$, *Planarity* $P = (\lambda_2 - \lambda_3)/\lambda_1$, *Sphericity* $S = \lambda_3/\lambda_1$, *Horizontality* $H = acos(\mathbf{v}_3 \cdot \mathbf{z})/||\mathbf{v}_3||$ and *Z range* $Z = Z_{max} - Z_{min}$, which are very common in the literature [15]. Calculation of Z range for each point is not limited to a sphere but to a vertical cylinder of a specific section (scale) around that point. In order to avoid the negative effect of outliers, Z coordinates are limited to an interval between the 5th and 95th percentiles.

28.2.2 Optimum Scale Estimation from Distance Correlation Functions

Distance correlation [12] is a measure of the degree of correlation, linear or non-linear, between two variables of arbitrary finite dimensions. When the data are ordered and close enough, it is possible to approximate distance correlation values for functions, and analyze them using methods for functional data. Particularly, we are interested in determining the global maximum of the distance correlation function, as it is supposed to be the point that captures the most relevant information concerning the relationship between the variables. A similar approach was used by the authors for variable selection in regression and classification problems [1, 6, 9]. $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ being two random vectors, distance correlation between X and Y is defined as

$$\mathcal{R}^2(X,Y) = \begin{cases} \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}, & \mathcal{V}^2(X)\mathcal{V}^2(Y) > 0 \\ 0, & \mathcal{V}^2(X)\mathcal{V}^2(Y) = 0 \end{cases} \tag{28.1}$$

where $\mathcal{V}^2(X,Y) = ||f_{X,Y} - f_X f_Y||^2 = \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(t,s) - f_X(t) f_Y(s)|^2}{|t|^{1+p}|s|^{1+q}}$ is the distance covariance, a measure of the distance between $f_{X,Y}$, the joint characteristic function of random vectors X and Y, and the product $f_X f_Y$ of the characteristics functions of X and Y, respectively. For their part, c_p and c_q are constants depending on the dimensions p and q, respectively.

Distance correlation has some advantages over other correlation coefficients, such as the Pearson correlation coefficient. First, it measures non-linear dependence. Second, X and Y do not need to be one dimensional variables. Third, $\mathcal{R}(X,Y) = 0 \Leftrightarrow X, Y$ are independent, that is, independence is a necessary and sufficient condition for the nullity of distance correlation.

Once the correlation distance has been determined for each feature at different scales $k \in \mathbb{R}$, we adjust a function $m : k \rightarrow \mathcal{R}(X,Y)$, and determine the values of k corresponding to global maximum of this function. Then, different supervised classification algorithms are applied using the features at those scales, and the results compared with those obtained when features are calculated at a specific number of scales at constant intervals or following an exponential function.

28.3 Experimental Results

28.3.1 Dataset

In order to evaluate the performance of the proposed methodology, we apply it to the Oakland 3D point cloud dataset [8], a benchmark dataset that has been previously used in different studies concerning point cloud segmentation and classification. The 3D point cloud was collected around the CMU campus in Oakland - Pittsburgh (USA) using a Mobile Laser Scanner (MLS), that consists of two-dimensional laser scanners, an Inertial Measurement Unit (IMU), and a Global Navigation Satellite system (GNSS), all of them calibrated and mounted on the Nablab 11 vehicle. Figure 28.1 shows a small part of the point cloud, where six labels have been marked.

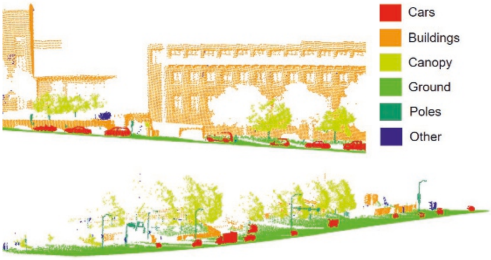


Fig. 28.1 Small area of the Oakland point cloud dataset. Each point has been assigned a label.

28.3.2 Neighborhood Selection

Consider a sample data $\{\mathbf{X}_i, \mathbf{G}_i\}_{i=1}^{n}$ where $\mathbf{X} = \left(X^1, ..., X^{J=5}\right)$, represents the vector of features (*linearity*, *planarity*, *sphericity*, *horizontality* and Z *range*), and $\mathbf{G} = (G_1, .., G_{m=5})$ the vector of classes (**cars**, **buildings**, **canopy**, **ground** and **poles**). For each sample i, each feature is evaluated at a regular grid of $N = 100$ scales measured in centimetres: $X_i^j = \left(X_i^j(t_1), X_i^j(t_2), \ldots, X_i^j(t_N)\right)$. Figure 28.2 shows a sample of $n = 150$ curves for each features registered in the interval $k \in [t_1 = 50, t_{100} = 300]$ and the corresponding functional mean, both colored by class label. Note the different performance of the features for the different classes and scales. For instance, *horizontality* takes high values for the **ground**, and it is uniform at different scales. However, this feature shows abrupt jumps at certain scales for the **poles**, that could correspond to edge effects. As expected, *linearity* takes high values for the **poles** and low values for the **buildings**.

Figure 28.3 shows the distance correlation functions for 100 repetitions of random samples of size $n = 750$ (150 per class), corresponding to each of the features extracted. A histogram of the global maximum of distance correlation curves for those repetitions is depicted at the bottom of the figure. As can be appreciated, most of the maxima (impact points) correspond to low scales, except for the Z *range* variable (5th - 95th range of z axis).

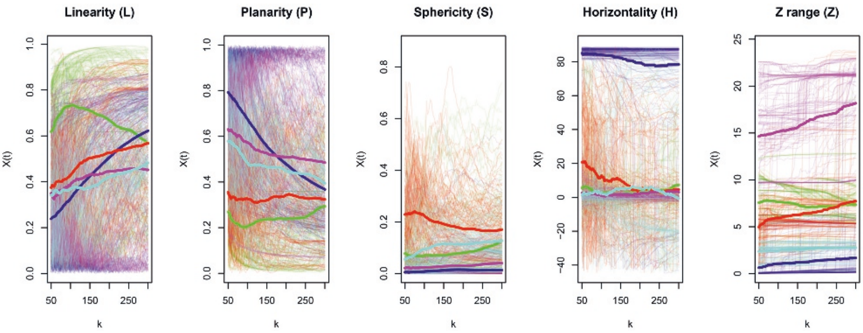


Fig. 28.2 A sample of features curves: poles (green), ground (blue), vegetation (red), buildings (magenta), and vehicles (cyan). Functional means for each class are represented as wider lines.

Our aim is to estimate an optimum neighborhood (scale) for each feature by means of distance correlation (DC), taking into account its advantage with respect to the Pearson coefficient. On the one hand, we calculated the distance correlation between the dependent variable (the label for each curve) and each of the features, see Figure 28.3. On the other hand, we calculated the distance correlation between the labels and two independent variables, *horizontality* and *Z range*, given that these features are more correlated with the dependent variable. In addition, DC between the dependent variable and the five features was also calculated. It is evident that DC functions are not very different for the five features analysed, and it is also evident that maximum values are reached at low scales, except for the range of z. In addition, it can also be appreciated that DC for *horizontality* and *Z range* are significantly

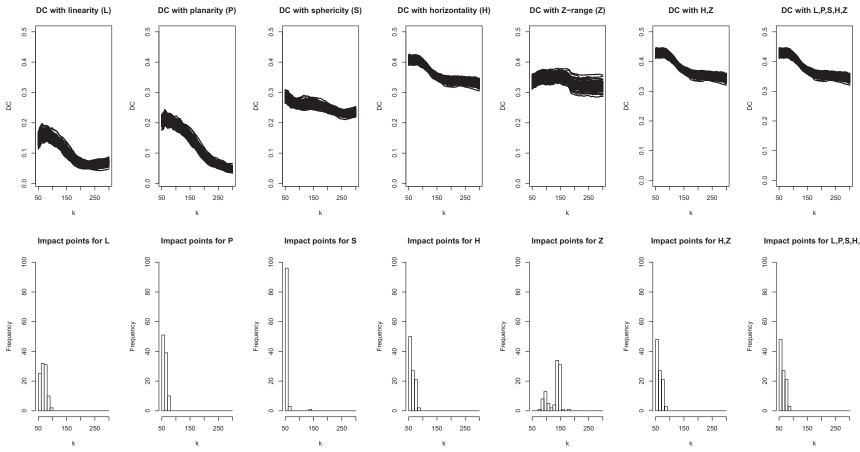


Fig. 28.3 Distance correlation functions between the group class and the features curves (top) and histogram with the scale for the global maximum on each function (bottom).

Table 28.1 Metrics of the classification using logistic regression (LR) and random forest classifier (RF) for different scenarios of k.

		Precision %					Recall %					F1 %				
Model	k	Poles	Grou.	Veg.	Build.	Cars	Poles	Grou.	Veg.	Build.	Cars	Poles	Grou.	Veg.	Build.	Cars
LR	50	70	96	74	**88**	76	74	**99**	72	77	82	72	98	73	82	78
LR	100	63	96	70	74	80	78	98	49	78	80	70	98	57	76	80
LR	175	51	96	66	64	75	63	93	32	78	82	57	95	43	70	78
LR	237	37	97	57	53	60	49	90	30	79	43	42	93	39	63	49
LR	300	28	95	49	49	46	34	88	32	80	25	31	91	39	61	31
LR	k_λ	61	97	70	74	77	73	96	50	78	80	66	97	75	75	52
LR	k_{mdc_2}	72	97	75	87	77	77	99	72	77	82	75	98	73	82	79
LR	k_{mdc_5}	74	97	75	**88**	78	76	**99**	71	79	87	75	98	73	**83**	**82**
RF	50	74	97	76	76	77	74	**99**	73	79	76	74	98	74	77	76
RF	112	**76**	**99**	**78**	71	**85**	**80**	96	74	79	77	**78**	98	76	78	81
RF	175	58	96	66	66	81	67	91	53	73	79	62	94	59	69	80
RF	237	52	96	57	67	76	63	89	45	68	79	57	92	50	67	77
RF	300	51	96	52	58	75	51	86	43	61	**89**	51	90	47	59	81
RF	k_λ	70	98	73	80	83	74	95	68	77	78	72	96	70	73	80
RF	k_{mdc_2}	**76**	98	**78**	74	82	77	98	77	**80**	77	76	98	77	77	78
RF	k_{mdc_5}	**76**	**99**	**78**	75	84	75	**99**	**80**	78	78	75	**99**	**78**	76	81

higher than for the other three features, which suggests that DC might be used not only to estimate the optimum scales but also to select the most important features to be included in the classification models.

In order to contrast the performance of this approach, we followed the proposal of [15] computing the optimal scale k_λ, corresponding to the minimum of the Shannon entropy E$_\lambda$, which depends on the normalized eigenvalues $e_i, i = 1, ..., 3$, of the local covariance matrix Σ: $E_\lambda = -e_1 ln(e_1) - e_2 ln(e_2) - e_3 ln(e_3)$ (2).

28.3.3 Classification

The scale corresponding to the most frequent values providing a global maximum (impact points) was used as input variable for two classification algorithms, multinomial logistic regression classifier (LR) and random forest classifier (RF) [14], in two scenarios: (a) k_{mdc_2}: only the features with the highest distance correlation values (*horizontality* and *Z range*) were included in the model and (b) k_{mdc_5}: all the features (*linearity*, *planarity*, *sphericity*, *horizontality* and *Z range*) were used to train the models. Additionally, we used the following values of the scale k: (c) k_λ, obtained according to equation (2), (d) k_{seq}, linearly spaced scales corresponding to the following values of k in centimetres (cm): $50, 112, 175, 237, 300$ and (e) k_{exp}, exponential spaced scales, that corresponds to $k = 1, 3, 7, 20, 55, 300$ cm. This last option arises from the fact that the global maximums of DC correspond to low scales.

Training data (150 per class) and test data (500 per class) were sampled from different areas of the point cloud, in order to ensure their independence. Table 28.1 shows the results of the classification for the test sample, in terms of precision, recall and F1-score, for each of the scales, using a logistic regression (LR) and random forest classifier (RF).

The metrics for the classification have quite different values depending on the category, see Table 28.1. Thus, the best results were obtained for **ground** class, followed by **cars**. Lower values were obtained for **poles**, **vegetation** and **buildings**. The results are very similar when the set of the five features (k_{mdc_5}), or just the two with the highest values of distance correlation (*horizontality* and *Z range*), are included in the model (k_{mdc_2}). In general, the models that use the scales corresponding to maximum distance correlation outperform the others, including that corresponding to the minimum of the Shannon entropy (k_λ), that did not turn out to be particularly good. Table 28.2 shows a decrease of the accuracy classification with the scale in both classifiers. So, it is better to calculate the scales using exponential function k_{exp} than using a linear function k_{seq}. This approach limits the number of scales to be calculated, thus reducing computing time.

Table 28.2 Total accuracy in % of the classification using logistic regression (LR) and random forest (RF) classifiers for sequential k_{seq}, exponential k_{exp}, k_λ, k_{mdc_2} and k_{mdc_5} scales.

	k_{seq}	50			100	150		200	250	300			
Model	k_{exp}	50	60	70	100		190			300	k_λ	k_{mdc_2}	k_{mdc_5}
LR		81	81	81	79	72	68	67	56	51	74	81	**83**
RF		80	81	82	81	75	72	71	66	65	78	81	81

28.4 Conclusions

Selecting optimum scales for supervised classification of 3D point clouds is relevant not only to improve the results but also to understand the effect of the features involved in the classification when the local neighborhood changes. We assume as hypothesis of our study that calculating the maximum of the distance correlation functions between the features (input variables) and the classes (output variable) can help to determine the optimum scale for classification and to select the most important variables at that scale. This hypothesis was tested on a benchmark 3D point cloud from an urban environment, and the analysis of the results indicates that our approach outperforms other common methods for scale selection, in particular one that uses specific scales at regular intervals and another that calculates the optimum scale using Shannon's information. Moreover, the analysis of the distance correlation functions for the different features provides information about the importance of these features in the classification. The best results were obtained when the five features, calculated at the optimum scale, were included in the classification model, but similar results were obtained when only the two features with the highest values of the correlation distance were considered. Accordingly, distance correlation function could be used as a filter for feature selection regardless of the classification algorithm. For future work, we plan to analyse a multi-scale analysis using significant structures of the features curves [3, 6].

Acknowledgements M. Oviedo acknowledges support from MTM2016-76969-P project (Spanish Ministry of Science, Innovation and Universities and the European Regional Development Fund).

References

[1] Berrendero, J.R., Cuevas, A., Torrecilla, J.L.: Variable selection in functional data classification: a maxima-hunting proposal. Statistica Sinica, 619–638 (2016)

[2] Brodu, N., Lague, D.: 3D terrestrial lidar data classification of complex natural scenes using a multi-scale dimensionality criterion: Applications in geomorphology. ISPRS Journal of Photogrammetry and Remote Sensing **68**, 121–134 (2012)

[3] Chaudhuri, P., Marron, J.: Scale space view of curve estimation. Ann. Stat, 408–428 (2000)

[4] Demantké, J., Mallet, C., David, N., Vallet, B.: Dimensionality based scale selection in 3d lidar point clouds (2011)

[5] Dittrich, A., Weinmann, M., Hinz, S.: Analytical and numerical investigations on the accuracy and robustness of geometric features extracted from 3D point cloud data. ISPRS Journal of Photogrammetry and Remote Sensing **126**, 195–208 (2017)

[6] Febrero-Bande, M., González-Manteiga, W., Oviedo de la Fuente, M.: Variable selection in functional additive regression models. Computational Statistics **34**(2), 469–487 (2019)

[7] Mallet, C., Bretar, F., Roux, M., Soergel, W., Heipke, Ch.: Relevance assessment of full-waveform lidar data for urban area classification. ISPRS Journal of Photogrammetry and Remote Sensing **66**(6), 571–584 (2011)

[8] Munoz, D., Bagnell, J., Vandapel, N., Hebert, M.: Contextual classification with functional max-margin Markov networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 975–982 (2009)

[9] Ordóñez, C., Oviedo de la Fuente, M., Roca-Pardiñas, J., Rodríguez-Pérez, J.L.: Determining optimum wavelengths for leaf water content estimation from reflectance: A distance correlation approach. Chemometrics and Intelligent Laboratory Systems **173**(15), 41–50 (2018)

[10] Pauly, M., Keiser, R., Gross, M.: Multi-scale feature extraction on pointsampled surfaces. Comput. Graph. Forum **22**(3), 281–289 (2003)

[11] Shannon, C.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948)

[12] Székely G.J., Rizzo, M.L., Bakirov, N.K.: Measuring and testing dependence by correlation of distances. Annals of Statistics **35**(6), 2769–2794 (2007)

[13] Ullman, S.: The Interpretation of Structure from Motion: Proceedings of the Royal Society of London. Series B, Biological Sciences **203**(1153), 405–426 (1979)

[14] Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S. Springer (2002)

[15] Weinmann, M., Jutzi, B., Hinz, S., Mallet, C.: Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers. ISPRS Journal of Photogrammetry and Remote Sensing **105**, 286–304 (2015)

Chapter 29
Generalized Functional Partially Linear Single-index Models

Mustapha Rachdi, Mohamed Alahiane, Idir Ouassou and Philippe Vieu

Abstract Single-index models are potentially important tools for multivariate nonparametric regression analysis. They generalize linear regression models by replacing the linear combination $\alpha_0^T X$ with a nonparametric component $\eta_0\left(\alpha_0^T X\right)$, where $\eta_0(\cdot)$ is an unknown univariate link function. [7] studied generalized partially linear single-index models (GPLSIM) where the systematic component of the model has a flexible semi-parametric form with a general link function. In this paper, we generalize these models to have a functional component, replacing the generalized partially linear single-index models $\eta_0\left(\alpha_0^T X\right) + \beta_0^T Z$ by $\eta_0\left(\alpha_0^T X\right) + \int_0^1 \beta_0^T(t)Z(t)\,dt$, where α is a vector in $\mathbb{R}^d$, $\eta_0(\cdot)$ and $\beta_0(\cdot)$ are unknown functions which are to be estimated. We propose estimators of the unknown parameter α_0 and the unknown functions $\beta_0(\cdot)$ and $\eta_0(\cdot)$ and we establish their asymptotic distributions. Then, we illustrate through some examples the models and the effectiveness of the proposed estimation methodology.

Mustapha Rachdi
Univ. Grenoble Alpes, AGEIS laboratory, UFR SHS, BP. 47, 38040 Grenoble Cedex 09, France,
e-mail: mustapha.rachdi@univ-grenoble-alpes.fr

Mohamed Alahiane (✉)
Université Cadi Ayyad, Ecole Nationale des Sciences Appliquées, Marrakech, Morocco,
e-mail: alahianemed@gmail.com

Idir Ouassou
Université Cadi Ayyad, Ecole Nationale des Sciences Appliquées, Marrakech and Université
Mohammed VI Polytechnique, 43140 Ben Guerir, Morocco, e-mail: i.ouassou@uca.ma

Philippe Vieu
Institut de Mathématiques de Toulouse, Université Paul Sabatier, 31062 Toulouse Cedex 9,
France. e-mail: vieu@math.univ-toulouse.fr

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_29

29.1 Introduction

Let H be an Hilbert space which is endowed with the scalar product $< \cdot, \cdot >_H$ and the norm $|| \cdot ||_H$. Let Y be a scalar response variable and $(X, Z) \in \mathbb{R}^d \times H$ be the predictor vector where $X = (X_1, \dots, X_d)$ and Z be a functional random variable which is valued in H.

For a fixed $(x, z) \in \mathbb{R}^d \times H$, we assume that the conditional density function of the response Y given $(X, Z) = (x, z)$ belongs to the following canonical exponential family

$$f_{Y|X=x,Z=z}(y) = \exp\Big(y\,\xi(x, z) - B(\xi(x, z)) + C(y)\Big), \tag{29.1}$$

where B and C are two known functions which are defined from $\mathbb{R}$ into $\mathbb{R}$, and $\xi \ : \ \mathbb{R}^d \times H \ \longrightarrow \ \mathbb{R}$ is the parameter in the generalized parametric linear model which is linked to the dependent variable

$$\mu(x, z) = \mathbb{E}\Big[Y|X = x, Z = z\Big] = B'(\xi(x, z)), \tag{29.2}$$

where B' denotes the first derivative of the function B.

In what follows we modelize $g(\mu(x, z))$ as a generalized functional partially linear single-index model by

$$g(\mu(x, z)) = \eta_0\left(\alpha^\top x\right) + \int_0^1 \beta(t)z(t)dt, \tag{29.3}$$

where $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_d) \in \mathbb{R}^d$ is the d-dimensional single-index coefficient vector, β is the coefficient function in the functional component, and η_0 is the unknown single-index link function which will be assumed to be sufficiently smooth.

29.2 Estimation Methodology

Let $(X_i, Y_i, Z_i)_{i=1,\dots,n}$ be a sequence of independent and identically distributed (i.i.d.) as (X, Y, Z) and, for each $i = 1, \dots, n$,

$$g\left(\mu\left(X_i, Z_i\right)\right) = \eta_0\left(\alpha^\top X_i\right) + \int_0^1 \beta(t)Z_i(t)\,dt. \tag{29.4}$$

We assume that the function η_0 is supported within the interval $[a, b]$ where $a = \inf(\alpha^\top X)$ and $b = \sup(\alpha^\top X)$.

We introduce a sequence of knots (k_m) in the interval $[a, b]$, with J interior knots, such that $k_{-r+1} = \cdots = k_{-1} = k_0 = a < k_1 < \cdots < k_J = k_{J+1} = \cdots = k_{J+r}$, where $J := J_n$ is a sequence of integers which increases with the sample size n.

Now, let $N_1 = J_n + r$ be the number of knots, $\left(B_j(u)\right)_{j=1,\dots,N_1}$ be the B-spline basis functions of order r, and $h = (b - a)/(J_n + 1)$ be the distance between the neighbors knots.

We introduce a new knots sequence $0 = t_0 < t_1 < \cdots < t_{k+1} = 1$ of $[0,1]$. Then, there exists $N_2 = k + r + 1$ functions in the B-splines basis which are normalized and of order r, such that $\beta(\cdot) \approx \delta^\top B_2(.)$ where $B_2(.) = \left(B_{21}(.), B_{22}(.), \ldots, B_{2N_2}(.)\right)^\top$ and $\delta \in \mathbb{R}^{N_2}$.

By setting

$$W\left(\int_0^1 Z(t)B_{21}(t)dt, \ldots, \int_0^1 Z(t)B_{2N_2}(t)dt\right), \tag{29.5}$$

and w and W_i are defined accordingly to (29.5), the mean function estimator $\widehat{\mu}(x,z)$ is then given by the evaluation of the parameter $\theta = (\alpha^\top, \gamma^\top, \delta^\top)^\top$ and by inverting the following equation $g\left(\widehat{\mu}(x,z)\right) B_{2j}(t)z(t)dt = \widehat{\gamma}^\top B_1\left(\widehat{\alpha}^\top x\right) + \widehat{\delta}^\top w$. Notice that the parameter $\theta = (\alpha^\top, \gamma^\top, \delta^\top)^\top$ is determined by maximizing the following quasi-likelihood rule

$$\widehat{\theta} = \left(\widehat{\alpha}^\top, \widehat{\gamma}^\top, \widehat{\delta}^\top\right)^\top = \underset{\theta=(\alpha,\gamma,\delta)\in\mathbb{R}^d\times\mathbb{R}^{N_1}\times\mathbb{R}^{N_2}}{\arg\max} l(\theta),$$

where

$$l(\theta) := l(\alpha,\gamma,\delta) = \frac{1}{n}\sum_{i=1}^n Q\left(g^{-1}(m_i), Y_i\right),$$

with

$$m(x,z) = \gamma^\top B_1\left(\alpha^\top x\right) + \delta^\top \langle z, B_2(.)\rangle$$

$$m_i := \gamma^\top B_1\left(\alpha^\top X_i\right) + \delta^\top W_i \quad \text{and} \quad m_{0i} = \gamma_0^\top B_1\left(U_{0i}\right) + \delta_0^\top W_i,$$

where $U_{0i} = \alpha_0^\top X_i$ with α_0, γ_0, δ_0, η_θ, β_0 denoting the true values, respectively, of α, γ, δ, η and β.

To overcome the constraint $\|\alpha\| = 1$ and $\alpha_1 > 0$ of the d-dimensional index α, we proceed by a re-parametrisation which is similar to [11]:

$$\alpha(\tau) = \left(\sqrt{1 - \|\tau\|^2}, \tau^\top\right)^\top \text{ for } \tau \in \mathbb{R}^{d-1}.$$

The true value τ_0, of τ, must satisfies $\|\tau_0\| \leq 1$. Then, we assume that $\|\tau_0\| < 1$. The jacobian matrix of $\alpha : \tau \longrightarrow \alpha(\tau)$ of dimension $d \times (d-1)$ is $J(\tau)$.

Notice that τ is unconstrained and is one dimension lower than α.

Finally, let

$$R(\tau) = \begin{pmatrix} J(\tau) & 0 \\ 0 & I_{N2\times N_2} \end{pmatrix}$$

the jacobian matrix of $(\alpha(\tau)^\top, \delta^\top)^\top$ which is of dimension $(d + N_2) \times (d + N_2 - 1)$.

Let

$$(\widetilde{\alpha}, \widetilde{\delta}) = \underset{\|\alpha\|_d=1,\ \delta}{\arg\max} \frac{1}{n}\sum_{i=1}^n Q\left(g^{-1}\left\{\widetilde{\eta}\left(\alpha^\top X_i\right) + \delta^\top W_i\right\}, Y_i\right)$$

Denote

$$m_i = \gamma^\top B_1\left(\alpha^\top X_i\right) + \delta^\top W_i, \quad T_i = \left(X_i^\top, W_i^\top\right)^\top,$$
$$m_{0i} = m_{0i}\left(X_i, W_i\right) = \gamma_0^\top B_1\left(\alpha_0^\top X_i\right) + \delta_0^\top W_i = \gamma_0^\top B_1\left(U_{0i}\right) + \delta_0^\top W_i \text{ with } U_{0i} = \alpha_0^\top X_i,$$
$$m_0(T) = \gamma_0^\top B_1\left(\alpha_0^\top X\right) + \delta_0^\top W = \gamma_0^\top B_1\left(U_0\right) + \delta_0^\top W \text{ with } U_0 = \alpha_0^\top X$$

and

$$(\widetilde{\tau}, \widetilde{\delta}) = \underset{\tau,\delta}{\arg\max}\ \widetilde{l}(\tau,\delta)$$

where

$$\widetilde{l}(\tau,\delta) = \frac{1}{n}\sum_{i=1}^{n} Q\left(g^{-1}\left\{\widetilde{\eta}\left(\alpha(\tau)^\top X_i\right) + \delta^\top W_i\right\}, Y_i\right)$$

Note that $\theta_\tau = (\tau^\top, \gamma^\top, \delta^\top)^\top$ is a $(d-1)\times N_1 \times N_2$-dimensional parameter, while θ is a $d \times N_1 \times N_2$-dimensional one.

Let

$$\rho_l(m) = \frac{1}{\sigma^2 V\left(g^{-1}(m)\right)}\left[\frac{d}{dm}\left(g^{-1}(m)\right)\right]^l$$

and denote

$$q_l(m,y) = \frac{\partial^l}{\partial m^l} Q\left(g^{-1}(m), y\right), \text{ for } l = 1, 2.$$

Then

$$q_1(m,y) = \left(y - g^{-1}(m)\right)\rho_1(m) \quad \text{and} \quad q_2(m,y) = \left(y - g^{-1}(m)\right)\rho_1'(m) - \rho_2(m).$$

So, $l\left(\theta_\tau\right)$ becomes

$$l\left(\theta_\tau\right) = \frac{1}{n}\sum_{i=1}^{n} Q\left(g^{-1}\left\{\gamma^\top B_1\left(\alpha^\top(\tau)X_i\right) + \delta^\top W_i\right\}, Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} Q\left(g^{-1}\left\{m_i\right\}, Y_i\right)$$

The score vector is then

$$S\left(\theta_\tau\right) = \frac{\partial l}{\partial \theta_\tau}\left(\theta_\tau\right) = \frac{1}{n}\sum_{i=1}^{n} q_1\left(m_i, Y_i\right)\xi_i(\tau,\gamma,\delta),$$

where

$$\xi_i(\tau,\gamma,\delta) = \begin{pmatrix} \gamma^\top B_1'\left(\alpha^\top(\tau)X_i\right) J^\top(\tau) X_i \\ B_1\left(\alpha^\top(\tau)X_i\right) \\ W_i \end{pmatrix}.$$

The Fisher Scoring update equations $\theta_\tau^{(k+1)} = \theta_\tau^{(k)} - \left[H\left(\theta_\tau^{(k)}\right)\right]^{-1} S\left(\theta_\tau^{(k)}\right)$, becomes

$$\theta_\tau^{(k+1)} = \theta_\tau^{(k)} + \left[\sum_{i=1}^{n} \rho_2\left(m_i^{(k)}\right)\xi_i\left(\tau^{(k)},\gamma^{(k)},\delta^{(k)}\right)\xi_i^\top\left(\tau^{(k)},\gamma^{(k)},\delta^{(k)}\right)\right]^{-1} \times\left[\sum_{i=1}^{n}\left(Y_i-\mu_i^{(k)}\right)\rho_1\left(m_i^{(k)}\right)\xi_i\left(\tau^{(k)},\gamma^{(k)},\delta^{(k)}\right)\right],$$

where $m_i^{(k)} = \gamma^{(k)\top}B_1\left(\alpha^{(k)\top}(\tau^{(k)})X_i\right) + \delta^{(k)\top}W_i$, for $1 \le i \le n$ and $\mu_i^{(k)} = g^{-1}\left(m_i^{(k)}\right)$.

It follows that

$$\widehat{\beta}(t) = \widehat{\delta}^\top B_2(t) \approx \delta^{(k)\top}B_2(t), \qquad \widehat{\eta}(t) = \widehat{\gamma}^\top B_1(t) \approx \gamma^{(k)\top}B_1(t),$$
$$\widehat{m}_i = \widehat{\gamma}^\top B_1\left(\alpha^\top(\widehat{\tau})X_i\right) + \widehat{\delta}^\top W_i \approx \gamma^{(k)\top}B_1\left(\alpha^\top\left(\tau^k\right)\right)X_i + \delta^{(k)\top}W_i,$$

where $\widehat{\mu}_i = g^{-1}\left(\widehat{m}_i\right)$, and $\widehat{\alpha} = \alpha\left(\tau^{(k)}\right)$ is the estimator of the single-index coefficient vector of the GFPLSIM model.

29.3 Assumptions

We present asymptotic properties of the estimators for the nonparametric components, the functional component, the single-index coefficient vector and the slope function of the GFPLSIM model. For this aim, we will need some assumptions.

Let φ, φ_1 and φ_2 be measurable functions on $[a,b]$. We define the empirical inner product $\langle\varphi_1,\varphi_2\rangle_n$ and its corresponding norm $\|\varphi\|_n$ as follows

$$\langle\varphi_1,\varphi_2\rangle_n = \frac{1}{n}\sum_{i=1}^{n}\varphi_1\left(U_i\right)\varphi_2\left(U_i\right) \quad\text{and}\quad \|\varphi\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}\varphi^2\left(U_i\right) \quad\text{where}\quad U_i = \alpha^\top X_i.$$

If φ, φ_1 and φ_2 are L^2-integrable, we define the theoretical inner product and its corresponding norm as follows

$$\langle\varphi_1,\varphi_2\rangle = \mathbb{E}\left[\varphi_1(U)\varphi_2(U)\right] \quad\text{and}\quad \|\varphi\|_2^2 = \mathbb{E}\left[\varphi^2(U)\right] = \int_a^b \varphi^2(u)f(u)du.$$

Let $v \in \mathbb{N}^*$ and $e \in (0,1]$ such that $p = v + e > 1.5$. We denote by $\mathcal{H}(p)$ the collection of functions g which are defined on $[a,b]$ whose v-th order derivative, $g^{(v)}$, exists and satisfies the following e-th order Lipschitz condition

$$\left|g^{(v)}\left(m'\right) - g^{(v)}(m)\right| \le C\left|m'-m\right|^e, \quad\text{for all}\quad a \le m, m' \le b.$$

(C1) The single-index link function $\eta_0 \in \mathcal{H}(p)$, where $\mathcal{H}(p)$ is defined as above.
(C2) For all $m \in \mathbb{R}$ and for all y in the range of the response variable Y, the function $q_2(m,y)$ is strictly negative and for $k = 1,2$, there exist some positive constants c_q

and C_q such that $c_q < \left|q_2^k(m,y)\right| < C_q$.
(C3) The marginal density function of $\alpha^\top X$ is continuous and bounded away from zero and is infinite on its support $[a,b]$.
(C4) For any vector τ, there exist positive constants c_τ and C_τ, such that

$$c_\tau I_{t\times t} \le \mathbb{E}\left[\left(\frac{1}{T}\right)\left(\frac{1}{T}\right)^\top \middle| \alpha^\top(\tau)X = \alpha^\top(\tau)x\right] \le C_\tau I_{t\times t},$$

where $t = 1 + N_1 + N_2$ and $T = (X^\top, W^\top)^\top$.
(C5) The number of knots N_n satisfies $n^{\frac{1}{2(p+1)}} \ll N_n \ll n^{\frac{1}{8}}$, for $p > 3$, where $N_n = N_1$.
(C6) The fourth order moment of the random variable Z is finite i.e., $\mathbb{E}\|Z(.)\|^4 \le C$ where C denotes a generic positive constant.
(C7) The covariance function $K(t,s) = \mathrm{Cov}(Z(t), Z(s))$ is positive definite.
(C8) The slope function β is a r-th order continuousely differentiable function i.e., $\beta \in C^r[0,1]$.
(C9) For some finite positive constants C_ρ, C_ρ^* and M_0

$$|\rho_1(m_0)| \le C_\rho \quad\text{and}\quad |\rho_1(m) - \rho_1(m_0)| \le |m - m_0| \;\text{ for all }\; |m - m_0| \le M_0.$$

(C10) For some finite positive constants C_g, C_g^* and M_1, the link function g, in the model (29.3), satisfies: $\left|\frac{d}{dm}g(m)\Big|_{m=m_0}\right| \le C_g$ and, for all $|m - m_0| \le M_1$,

$$\left|\frac{d}{dm}g^{-1}(m) - \frac{d}{dm}g^{-1}(m)\Big|_{m=m_0}\right| \le C_g^* \left|m - m_0\right|.$$

(C11) There exists a positive constant C_0, such that

$$\mathbb{E}(\epsilon^2 | U_{\tau,0}) \le C_0, \text{ where } \epsilon = Y - g^{-1}\left(m_0(T)\right).$$

29.4 Some Asymptotics

Next we formulate several assertions on the considered estimators.

Estimation of the nonparametric component. The following theorem states the convergence, with rates, of the estimator $\widehat{\eta}$.

Theorem Under assumptions (C1)–(C8), we have

$$\|\widehat{\eta} - \eta_0\|_2 = O_p\left\{\sqrt{N_n}\left(\frac{1}{\sqrt{n}h} + h^p\right)\right\} \quad\text{and}\quad \|\widehat{\eta} - \eta_0\|_n = O_p\left\{\sqrt{N_n}\left(\frac{1}{\sqrt{n}h} + h^p\right)\right\},$$

where O_p denotes a “grand O of Landau" in probability. □

Estimation of the slope function.

Theorem Under assumptions (C1)–(C8), we have

$$\|\widehat{\beta}(\cdot) - \beta_0(\cdot)\|^2 = O_p\left(N_n\left(h^p + \frac{1}{\sqrt{nh}}\right)^2\right).$$

Estimation of the parametric components. The next theorem shows that the maximum quasi-likelihood estimator is root-n consistent and is asymptotically normal, although the convergence rate of the nonparametric component $\widehat{\eta}$ is slower than root-n. Before enouncing the theorem, let us denote by

$$\Upsilon\left(u_{\tau,0}\right) = \frac{\mathbb{E}\left[X\rho_2\left(m_0(T)\right)\Big|U_{\tau,0} = u_{\tau,0}\right]}{\mathbb{E}\left[\rho_2\left(m_0(T)\right)\Big|U_{\tau,0} = u_{\tau,0}\right]} \quad , \quad \Gamma\left(u_{\tau,0}\right) = \frac{\mathbb{E}\left[W\rho_2\left(m_0(T)\right)\Big|U_{\tau,0} = u_{\tau,0}\right]}{\mathbb{E}\left[\rho_2\left(m_0(T)\right)\Big|U_{\tau,0} = u_{\tau,0}\right]},$$

$$\Phi(x) = \Phi\left(U_{\tau,0}, x\right) = x - \Upsilon\left(u_{\tau,0}\right) \text{ and } \Psi(w) = \Psi\left(U_{\tau,0}, w\right) = w - \Gamma\left(u_{\tau,0}\right).$$

Theorem Under assumptions (C1)–(C8), the constrained quasi-likelihood estimators $\widehat{\alpha}$ and $\widehat{\delta}$ with $\|\widehat{\alpha}\|_d = 1$ are jointly asymptotically normally distributed, i.e.,

$$\sqrt{n}\begin{pmatrix}\widehat{\alpha} - \alpha_0\\ \widehat{\delta} - \delta_0\end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, R\left(\tau_0\right) D^{-1} R^\top\left(\tau_0\right)\right),$$

where $\xrightarrow{\mathcal{D}}$ denotes the convergence in distribution, and

$$D = \mathbb{E}\left[\rho_2\left(m_0(T)\right)\begin{pmatrix}\eta_0'\left(U_{\tau,0}\right) J^\top\left(\tau_0\right)\Phi(X)\\ \Psi(W)\end{pmatrix}\begin{pmatrix}\eta_0'\left(U_{\tau,0}\right) J^\top\left(\tau_0\right)\Phi(X)\\ \Psi(W)\end{pmatrix}^\top\right],$$

and

$$R(\tau) = \begin{pmatrix}J(\tau) & 0\\ 0 & I_{N_2\times N_2}\end{pmatrix}.$$

Comments on the assumptions. The smoothnes condition in **(C1)** describes that the single-index function $\eta_0(\cdot)$ can be approximated by functions in the B-spline space with a normalized basis. On the other hand, the condition **(C2)** ensures the uniqueness of the solution, whereas condition **(C3)** is a smoothness assumption of the joint and the marginal density functions of $\alpha^\top X$ and X. On the other hand, condition **(C5)** allows to obtain the rate of growth of the dimension of the spline space with respect to the sample size. Conditions **(C6)** and **(C7)** are required on the covariates function Z and **(C8)** is a smoothness assumption on the slope function. Conditions **(C4)**, **(C9)**, **(C10)** and **(C11)** are a Lemma's technical assumptions.

29.5 A Numerical Study

We conducted (i) a simulation study and (ii) an application on some real datasets. The obtained results are very satisfactory and very promising. In order to save space, we cannot present it here. In this practical study, all the parameters have been chosen with precision and all the procedures are well controlled. These results can be requested from the authors. The proofs will be published in an extended version of this paper.

References

[1] Caroll, R.J., Fan, J., Gijbels, I., Wand, M.P.: Generalized partially linear single-index models. J. Amer. Statist. Assoc. **92**(438), 477-489 (1997)

[2] De Boor, C.: A practical guide to splines, revised Edition, Vol. 27 of Applied Mathematical Sciences. Springer-Verlag, Berlin (2001)

[3] Huang, J.: Efficient estimation of the partly linear additive Cox model. Ann. Statist. **27**(5), 1536–1563 (1999)

[4] Pollard, D.: Asymptotics for least absolute deviation regression estimators. Econometric Theory **7**, 186–199 (1991)

[5] Stone, C.J.: The dimensionality reduction principle for generalized additive models. Ann. Statist. **14**(2), 590–606 (1986)

[6] Van der Vaart, A.W., Wellner, J.A.: Weak convergence and empirical processes with applications to statistics. Springer, New-York (1996)

[7] Wang, L., Cao, G.: Efficient estimation for generalized partially linear single-index models. Bernoulli **24**(2), 1101–1127 (2018)

[8] Wang, L., Yang, L.: Spline estimation of single-index models. Statistica **19**, 765–783 (2009)

[9] Xue, L., Yang, L.: Additive coefficient modeling via polynomial spline. Statist. Sinica **16**, 1423–1446 (2006)

[10] Yu, P., Du, J., Zhang, Z.: Single-index partially functional linear regression model. Statist. Papers (2018)

[11] Yu, Y., Ruppert, D.: Penalized Spline Estimation for Partially Linear Single-Index Models. J. Amer. Statist. Assoc. **97**(460), 1042–1054 (2002)

Chapter 30
Functional Outlier Detection through Probabilistic Modelling

Álvaro Rollón de Pinedo, Mathieu Couplet, Nathalie Marie, Amandine Marrel, Elsa Merle-Lucotte and Roman Sueur

Abstract
Functional data consist in the most typical case of one-dimensional curves that represent the evolution of some physical parameter of interest with time. However, the analysis of this kind of objects is far from being simple, and the possibility of treating contaminated data is a classical problem that can arise in this framework as frequently as in the multivariate one. This justifies the development a new functional outlier detection technique based on functional measures capable of capturing the the outlyingness in the magnitude and shape sense that is presented in this paper.

30.1 Introduction

In recent times, the increasing use of computer codes in simulation studies, as well as the generalization of the use of sensors have resulted in the generation of large quantities of high-dimensional data which in many cases can be considered to be functional. It is the case for instance in the domain of thermal-hydraulics simulation [9] or the registration of the concentration of atmospheric pollutants [6].

Álvaro Rollón de Pinedo (✉)
EDF R&D, 6 quai Watier, 78400 CHATOU, France, e-mail: alvaro.rollon-de-pinedo@edf.fr

Mathieu Couplet
EDF R&D, 6 quai Watier, 78400 CHATOU, France, e-mail: mathieu.couplet@edf.fr

Nathalie Marie
CEA Cadarache, 13108 Saint-Paul-lez-Durance, France, e-mail: nathalie.marie@cea.fr

Amandine Marrel
CEA Cadarache, 13108 Saint-Paul-lez-Durance, France, e-mail: amandine.marrel@cea.fr

Elsa Merle-Lucotte
LPSC Grenoble, 53 avenue des Martyrs, France, e-mail: merle@lpsc.in2p3.fr

Roman Sueur
EDF R&D, 6 quai Watier, 78400 CHATOU, France, e-mail: roman.sueur@edf.fr

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_30

The domain of functional data analysis is relatively recent, and the main references on the field are the works of Ramsay and Silverman [12] and Ferraty [8]. These data will be supposed to belong to an infinite-dimentional vector space [3] and, more specifically, in the context of these works, we shall consider functional random variables.

Given a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, where Ω is the sample space, $\mathcal{A}$ is the event space, and $\mathbb{P}$ is a probability measure, as well as a certain functional space $\mathcal{F}$, then a random variable is called *functional*, if it takes its values in a vector space of infinite dimension $(\mathcal{F}, \mathcal{B}_{\mathcal{F}})$, where $\mathcal{B}_{\mathcal{F}}$ is the Borel σ-algebra of the space. It is then a measurable application $X : \Omega \rightarrow \mathcal{F}$.

Taking this definition into consideration, the considered functions could have an arbitrary number of dimensions, but in our context we shall focus on one-dimensional functions (curves). This is the classical case of the evolution of a physical parameter as a function of time.

In this context, the development of tools that allow the detection of *anomalous* or *outlying* curves in certain sets of functional data is not obvious, but can have a major impact on the conclusions that are extracted from the study of these data. An identification tool can be used in order to automatically identify curves associated with unexpected physical phenomena, measurement errors or other relevant pieces of information needed by the users before extracting conclusions from the original set of data.

In the following parts of the document we will expose the main points regarding the domain of functional outlier detection (Sections 30.2.1 and 30.2.2), the new procedure that has been implemented (Sections 30.2.3 and 30.2.4), as well as some mathematical properties and a test through known theoretical examples (Sections 30.2.5 and 30.3).

30.2 Functional Outlier Detection

30.2.1 Introductory Aspects

The first main remark that should be made is that there is no formal definition of what constitutes an outlier in a set of data. This is the reason why most references prefer not to give a precise definition, and mostly define them as data within the studied set that behave in an abnormal way when compared to the majority of objects [2].

As it is usual, the field of outlier detection is far more developed in the multivariate context than in the functional one. Most existing detection techniques in the multivariate framework rely on some notion of density or distance between the objects of the set of data, but this approach is not necessarily adapted for functional data. As an example, in a set of periodic curves, a particular one could be considered to have an outlying nature because its frequency is significantly different from the others, even though that curve could be placed in a highly dense region. This is the reason why some authors [4] make an essential distinction between magnitude (those who deviate from the normal set by a certain distance criteria), and shape (those whose form is significantly different from most of the curves of a set) functional outliers.

In [10], a general classification of the main existing methods of functional outlier detection is presented. In general, they can be divided into three main groups: (i) Two stage approaches, where there is a dimensionality reduction step that allows the application of a classical multivariate methods, (ii) Non-parametric approaches which do not intend to represent the data in a space of lower dimension, but rather quantify the differences between the curves through similarity measures, and finally, (iii) model-based approaches, which aim to model the probability density function (pdf) of some scalar descriptor of the functional data.

Our method follows the line of the last possibilities, and is based on two notions. The first one is the fact that in all generality, it is true that both magnitude and shape outliers can be of interest, and therefore there is a need for a method capable of capturing both notions of outlyingness and the use of non-parametric functional measures adapted to these characteristics seems logical. On top of that, when complex data are analysed and the notions of orientation of curves or shape become relevant, it is convenient to adjust some kind of probabilistic model over the random variables of interest (such as the aforementioned non-parametric descriptors), so that data objects that fit the distribution should have high probabilities of occurence.

30.2.2 Functional Measures

Two functional measures (this is a slight abuse of terminology, since the second is not strictly a measure) are retained thanks to their sensitivity to magnitude and shape differences. They are the *h-modal* depth [11] (also called *h-mode* depth), a definition of local depth due to the fact that it does not take into account the whole empirical sample of functional data, but only a slightly smaller window in order to guarantee to a certain degree that multimodality distributions can be detected, and the Dynamic Time Warping (DTW) [15], which is a measure of correspondence between temporal sequences that allows the comparison of their shapes.

The *h-mode* depth of a realization $z \in C([0,1])$ with respect to $Z \sim P \in \mathcal{P}(C([0,1]))$ is defined as:

$$h_M(z;P) = \mathbb{E}\Big(\frac{1}{h(P)}K\Big(\frac{\|z-Z\|}{h(P)}\Big)\Big), \tag{30.1}$$

where $K : \mathbb{R} \longrightarrow \mathbb{R}$ is a kernel function and $\|\cdot\|$ is a chosen norm on the considered functional space. There are several kernel functions available, but one that is typically used is the Gaussian Kernel:

$$K(t) = \frac{2}{\sqrt{2\pi}}\exp\Big(-\frac{t^2}{2}\Big), t > 0$$

and h represents the bandwidth, which is usually taken as the 15th percentile of the empirical distribution of $\|z_i - z_k\|; i, k = \{1, ..., N\}$ [7].

The other main measure that will be used mainly for shape outlier detection, comes from the time series domain, the Dynamic Time Warping (DTW), which provides a measure of similarity and correspondence between two sequences of data.

Given two sequences $X := (x_1, x_2, ..., x_N); N \in \mathbb{N}$ and $Y := (y_1, y_2, ..., y_M); M \in \mathbb{N}$, as well as a feature space $\mathcal{S}$. Then, $x_n, y_m \in \mathcal{S}$ for $n \in [1 : N]$ and $m \in [1 : M]$, and we can define a local cost measure (sometimes also called local distance measure), which is a function: $c : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$.

In this case, an (N, M)-warping path, is a sequence $p = (p_1, ..., p_L)$ with $p_l = (n_l, m_l) \in [1 : N] \times [1 : M]; \forall l \in [1 : L]$ which also satisfies the following conditions:

- Boundary condition: $p_1 = (1, 1)$ and $p_L = (N, M)$.
- Monotonicity condition: $n_1 \leq n_2 \leq ... \leq n_L$ and $m_1 \leq m_2 \leq ... \leq m_L$.
- Step size condition: $p_{l+1} - p_l \in \{(1, 0), (0, 1), (1, 1)\}$ for $l \in [1 : L - 1]$.

The total cost $c_p(X, Y)$ of a given warping path is:

$$c_p(X, Y) := \sum_{l=1}^{L} c(x_{n_l}, y_{n_l}).$$

Finally, an *optimal warping path* between X and Y is a warping path p^* having minimal total cost among all possible warping paths. Having defined this path, *DTW distance* $DTW(X, Y)$ between X and Y is defined as the total cost of the optimal warping path. This way, if two curves are **normalised**, it is possible to compare the shape of curves by making use of modified versions of this DTW measure.In order to simplify the notation, from here onwards **both notions will be expressed as**:

$$h_M(z_i, P) \equiv d_{i,1}(z_i), \forall z \in (\mathbb{R}^Q)^{\mathcal{U}},$$
$$DTW(z_i, z_j) \equiv d_{i,j,2}(z_i, z_j), \forall z_i, z_j \in (\mathbb{R}^Q)^{\mathcal{U}},$$
$$\text{and } \overline{DTW(z_i)} = d_{i,2} = \frac{1}{N} \sum_{j=1}^{N} d_{i,j,2}(z_i, z_j).$$

30.2.3 Methodology

The association of a value of depth for each functional datum z_i, and of DTW for each pair of data z_i, z_j is the basis of this functional outlier detection technique. Since the analysed curves will be considered functional random variables, any real-valued measure applied to them shall be treated as a real-valued random variable. In this case, the h-mode depth and the DTW shall be denoted $\{D_1, D_2\}$.

The main idea of the algorithm is to adjust a probabilistic model such as the Gaussian Mixture Model (GMM) to the two real-valued random variables that are supposed to quantify the degree of outlyingness in the magnitude and shape sense, i.e., the variables D_1, D_2, which will have a set of realizations depending on the original data: $\{d_{i,1}, d_{i,2}\}_{i=1}^{N}$.

Following this logic, a joint bivariate parametric probabilistic model to the pair of data $\{d_{i,1}, d_{i,2}\}$ for every functional datum z_i. By proceeding this way, a particular curve can be considered as an outlier even if it does not have an extreme value with regard to both measures if they are looked upon independently, but the combination of its depth and DTW values can make it have an outlying nature.

The form of the GMM model is:

$$\hat{f}(d_1, d_2) = \sum_{k=1}^{K} \hat{\alpha}_k \mathcal{N}(\hat{\mu}_k, \hat{\Sigma}_k),$$

where $k \in \{1, .., K\}$ is the number of bivariate Gaussian probability functions of mean vector $\hat{\mu}_k$ and covariance matrix $\hat{\Sigma}_k$ and the weight vector $\{\alpha_1, ..., \alpha_K\}$ verifies $\sum_{k=1}^{K} \alpha_k = 1$ and $\alpha_k \geq 0$. The function $\hat{f}(d_1, d_2)$ is the joint probability density function of the random variables D_1, D_2, which associates a depth and a DTW value to each functional realization. The estimation of all these parameters can be done through the Expectation Maximization (EM) algorithm [5].

Once the joint PDF is adjusted, it is possible to extract probabilistic conclusions on the random variables that are modelled. For instance, if a certain set of pairs of values of $\{d_{i,1}, d_{i,2}\}$ defines a closed region such that a certain percentage of probability mass, Q of the PDF is kept, there is a notion of multivariate quantile that can be interpreted. More formally, if there exists a value q such that:

$$\int_{\mathbb{R}\times\mathbb{R}} \mathbb{1}_{\{\hat{f}\,(d_1,d_2)\leq q\}} = Q,$$

Then the identity $\hat{f}(d_1, d_2) = q$ defines a (closed, due to the concave nature of the GMM) curve that can be written as $\gamma = g(d_1, d_2) \equiv \partial D$, where D is the open domain in $\mathbb{R} \times \mathbb{R}$ with frontier defined by ∂D, and such that $\forall (d_{i,1}, d_{i,2}) \notin D$, z_i is considered to be an **outlier**.

Naturally, the estimation of the joint PDF is dependent on the estimated set of parameters $\{\hat{\alpha}_k, \hat{\mu}_k, \hat{\Sigma}_k\}_{k=1}^{K}$, which is dependent on the quantity of available functional data.

30.2.4 Estimation of the Probability of Being an Outlier

A practical way of implementing this procedure is to generate a bootstrap sample from the original curves and perform an estimation of $\gamma, \hat{\gamma}(d_1, d_2) = \gamma_{\{\hat{\alpha}_k, \hat{\mu}_k, \hat{\Sigma}_k\}_{k=1}^{K}}$ for each group. By proceeding this way, it could be possible to estimate confidence intervals for the estimated curve, or to associate a score to the likelihood of being an outlier.

In summary, and taking the original curves $\{z_i\}_{i=1}^{N}$ as starting point, several bootstrap (sampling from the curves with replacement) groups will be obtained, each one with the same number of samples from $\{z_i\}_{i=1}^{N}$. Then an independent

GMM model will be fitted to the data of each group through the EM algorithm, and the desired quantile curve γ will be estimated for every group. By doing this, it is possible to quantify if every curve is considered as outlier by fulfilling the criterion of each bootstrap group.

More formally, if we have defined a certain number of groups B, indexed by b such that $b \in \{1, ..., B\}$, then for each functional datum $z_i, i \in \{1, ..., N\}$ it is possible to define a binary random variable W such that:

$$W = \begin{cases} 1 \; if \;\; z_i \notin D_b \\ \\ 0 \; if \;\; z_i \in D_b \end{cases}$$

where D_b is the previously mentioned acceptance region (open domain) in $\mathbb{R} \times \mathbb{R}$ for the bootstrap group b. Then, by computing the expectancy of the realization of the random variable $W_i = W(z_i)$ for each z_i in each bootstrap group, it is possible to quantify a score that will be equivalent to the probability of being considered an outlier according to the aforementioned measures $\{d_{i,1}, d_{i,2}\}$.

As an example, the value $\mathbb{P}(W_i = 1)$ quantifies the probability that the functional datum z_i is an outlier according to our criteria.

30.2.5 Estimation of the Number of Outliers

It is possible to perform an estimation of the number of outliers coherently with the detection technique. According to this, if μ is the actual percentage of outliers, the estimation of the number of outliers would be:

$$\hat{\mu} = \frac{\sum_{i=1}^{N} \mathbb{1}_{(d_{i,1}, d_{i,2}) \notin D_i}}{N},$$

In this case, logically, the expectation of the estimator is equal to the significance value of the rejection region defined by the GMM model if the sample is formed by independent and identically distributed data :

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}\big[\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{(d_{i,1}, d_{i,2}) \notin D_i}\big] = \frac{1}{N} \cdot N \cdot \mathbb{P}[(d_1, d_2) \notin D] = \mathbb{P}[(d_1, d_2) \notin D] = \alpha$$

And so the bias of the chosen estimator can be quantified as:

$$\mathbb{E}[\hat{\mu} - \mu] = \mathbb{E}[\hat{\mu}] - \mathbb{E}[\mu] = \alpha - \mu.$$

30.3 Confrontation to Theoretical Test Case

Two models will be retained in order to test the detection capacities of the algorithm. The first one constitutes magnitude and shape outlier that does not deviate greatly with respect to the other curves and the second is a shape outlier. The models are:

- The first model is the funtion generator: $f_1(t) = sin(t) + 0.2cos(100t) + u$ where the variable u is a realization of the distribution $\mathcal{U}[-1, 1]$. The outliers are generated through $o(t) = sin(0.3t) + n$ where n is a realization of the distribution $N(0, 1)$.
- The second model that generates the functions is $f_2(t) = 4t + e$ where e is a Gaussian Process of mean $\mu = 0$ and correlation function given by $\gamma(t_1, t_2) = 0.3e^{-\frac{|t_1-t_2|}{0.3}}$, whereas the outliers follow the distribution: $o_2(t) = 4t$.

For every test, 50 main curves and 1 outlier were generated.

The outlier detection technique was applied to both cases 200 times in order to estimate the false positive detection rate (percentage of data identified as outliers even though they are not), as well as the detection rate (percentage of correctly identified outliers). Both of them can me measured depending on how high the estimated probability of being an outlier must be. For instance, it is possible to impose a score outlyingness of 50% or 75% for a datum to be considered as outlier. The results for these two models are compared with those of the *functional boxplot* [14] and *directional outlyingness* [13](in the last case, by taking the most extreme values of its univariate boxplot).

Table 30.1 Detection rates for the $\mathbb{E}[W_i] \geq 0.5$ and $\mathbb{E}[W_i] \geq 0.75$ criteria

50%	Model 1	Model 2	**75%**	Model 1	Model 2
Av. Detection rate	100%	100%	Av. Detection rate	100%	78%
False Positive Rate	0.22%	1.5%	False Positive Rate	0.17%	0.83%

Table 30.2 Detection rates for Directional Outlyingness and Functional Boxplot

DO	Model 1	Model 2	**FB**	Model 1	Model 2
Av. Detection rate	100%	64%	Av. Detection rate	42.6%	0%
False Positive Rate	5.36%	9.8%	False Positive Rate	0%	0.4%

30.4 Conclusions

In this work a novel functional outlier detection technique has been presented and several results are exposed. Its outlier detection rates are competitive against state of the art detection algorithms and it presents several other advantages.

One of them is the fact that not only a binary indicator of outlyingness is provided, but also to what degree each functional datum can be considered to be an outlier. This allows the user to establish a prioritization when analysing the underlying reasons for this outlyingness behind the data. Another important advantage is the possibility of distinguishing the reasons why a datum is regarded as an outlier (if it is considered as such in the magnitude or shape sense), by analysing the marginal distributions of the fitted probabilistic model.

References

[1] Arribas-Gil, A., Romo, J.: Shape outlier detection and visualization for functional data: the outliergram. Biostatistics **15**, 603–619 (2014)

[2] Barreyre, C., Laurent, B., Loubes, J.-M., Cabon, B., Boussouf, L.: Multiple testing for outlier detection in functional data. arXiv:1712.04775 (2017)

[3] Chagny, G.: Statistique pour données fonctionnelles. Université Paris-Dauphine (2016)

[4] Dai, W., Mrkvicka, D., Sun, Y., Genton, M.G.: Functional Outlier Detection and Taxonomy by Sequential Transformations. arXiv: 1808.05414 (2018)

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

[6] Febrero-Bande, M., Galeano, P., González-Manteiga, W.: Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. Environmetrics **19**, 331–345 (2008)

[7] Febrero-Bande, M., Galeano, P., González-Manteiga, W.: Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. Environmetrics **19**, 331–345 (2008)

[8] Ferraty, F.: Modélisation Statistique pour Variables Aléatoires Fonctionnelles : Théorie et Applications. Habilitation à diriger des recherches, Université Paul Sabatier (2003)

[9] Iooss, B., Marrel, A.: Advanced Methodology for Uncertainty Propagation in Computer Experiments with Large Number of Inputs. Nuclear Technology **205**(12), 1588–1606 (2019)

[10] Jacques, J., Preda, C.: Functional data clustering: a survey. Advances in Data Analysis and Classification. Springer Verlag (2014)

[11] Nagy, S.: Consistency of h-mode depth. Journal of Statistical Planning and Inference **165**, 91–103 (2015)

[12] Ramsay, S.O., Silverman, B.W.: Functional Data Analysis. Springer Series in Statistics. Springer (2005)

[13] Sun, Y., Genton, M.G.: Directional outlyingness for multivariate functional data. Computational Statistics & Data Analysis **101**, 50–65 (2011)

[14] Sun, Y., Genton, M.G.: Functional boxplots. Journal of Computationl and Graphical Statistics **20**(2), 316-334 (2011)

[15] Zhang, Z., Tavenard, R., Bailly, A., Tang, X., Tang P., Corpetti, T.: Dynamic Time Warping Under Limited Warping Path Length. Information Sciences **393**, 91–107 (2017)

Chapter 31
Topological Object Data Analysis Methods with an Application to Medical Imaging

Chen Shen and Vic Patrangenaru

Abstract We apply ideas from algebraic topology to study distributions on object spaces. We present a framework for using persistence landscapes to vectorize persistence diagrams as in Bubenik(2015)[3] and Patrangenaru et al.(2018)[13]. We apply these methods to analyze data from The Cancer Imaging Archive (TCIA), using a technique developed earlier for regular types of digital images. The aim of this study is a comparison of brain images of CPTAC Glioblastoma patients with similar images from clinically normal individuals. Result shows persistence landscape may capture topological features distinguishing the two groups.

31.1 Introduction

Glioblastoma multiforme (GBM), is the most aggressive cancer that begins within the brain (see Holland(2000)[11]). We apply topological data analysis(TDA) methods to object data distributions extracted from CT scans, to detect GBM. TDA provides ways of analyzing features of object distributions that could not be addressed by standard object data analysis (ODA). Objects are usually regarded as points on a manifold or more generally on a stratified space, called *object space*,whose structure is highly nonlinear, making significant standard statistical analysis obsolete. While standard ODA methods are usually addressing location and spread parameter data analysis, TDA delivers answers regarding distributions on the object space on the whole, as they are based on Algebraic Topology (AT) invariants. AT helps detecting “holes", “voids", “tubes" or other missing components in topological spaces via *homology invariants*. A multi-scale summary of such invariants is provided by applying persistence topological invariants on metrizable distance spaces. TDA methods are quite flexible, and there are many possible ways to apply them to object data. In this

Chen Shen
Florida State University, U.S.A., e-mail: cs15j@my.fsu.edu

Vic Patrangenaru (✉)
Florida State University, U.S.A., e-mail: vpatrangenaru@fsu.edu

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_31

paper, we will use *persistent homology*. The crucial step is to encode the object data through an *increasing filtration* of the support of a distribution on the object space, containing enough structure so that the subsequent statistical analysis would be successful. Statistics is rooted in the theory of errors; often times a continuous random variable, regarded as a combination of measurements involving random errors, is described in terms of its p.d.f., as a mixture of small'bumps" around certain relevant locations. Such probability models for error prone measurements, say for example, multivariate normal distributions, Dirichlet distributions, Wishart distributions and their mixtures, are models that can not capture the above mentioned topological features of data. In TDA, one considers a sample of observation from a random object, and apply to it persistent homology algorithms. We build an increasing family of *simplicial complexes*, called *Vietoris-Rips complexes*, from the pairwise distances between the point in the sample. We calculate their persistent homology and use *death vectors* and *persistence landscapes* to vectorize the data, which allows applying functional data analysis techniques. The landscapes, that constitute a functional summary of the object data, substantially simplify our statistical analysis.
A brief outline is given as follow. In section 2 we list the main steps of our topological data analysis. We define death vectors and persistent landscapes (PLs) and introduce metric structures of PLs; this allows us to apply functional data analysis techniques to set statistical hypotheses and construct statistic tests. Section 3 is dedicated to the brain cancer image data analysis based on The Cancer Imaging Archive(TCIA), where TDA methods are used.

31.2 Object TDA via persistence homology

Due to space limitations, for basic AT definitions needed here, we send the Statistics readership to Section 3.4. in Patrangenaru and Ellingson(2015)[12]. TDA summarizes the topological and geometric structure of data by applying tools from AT to certain geometric structures built from the data on hand. Suppose we have points on a manifold, topology studies the connectivity of these points. We could replace each point by a disk(ball) with radius ε. As the radius increases, points will be connected with edges and so the number of connected components will decrease. The connected components are considered as 0 degree topological features, and the number of connected components is denoted as the 0-th Betti number, β_0. While connecting these points with edges, simplices will also be created to produce topological features in higher dimensions. A 0-simplex is a single point or vertex, a 1-simplex is the line segment or edge determined by 2 distinct vertices, 2-simplex is the solid triangle determined by 3 vertices, that do not lie on a line, and so on. The k-th Betti number β_k counts the number of k-degree topological features. A set of simplices whose vertices all have pairwise nonempty intersections is called the Vietoris-Rips (VR) complex. In general, there is no single proximity parameter ε that yields a Vietoris-Rips complex $\mathcal{R}_\varepsilon$ which best describes the topological and geometric structure from which that data point cloud was sampled. Instead one considers all possible values of ε and one determines which topological features persist as ε increases. Persistent homology completely describes how homology persists as one steps through the filtration (see Edelsbruner et al.(2002)[9]). For example, consider a filtration of Vietoris-Rips complexes $\mathcal{R}_{\varepsilon_0} \subset \mathcal{R}_{\varepsilon_1} \subset \cdots \subset \mathcal{R}_{\varepsilon_m}$ for $\varepsilon_0 < \varepsilon_1 < \cdots < \varepsilon_m$. One is interested in topological features that persist as the proximity parameter ε ranges from ε_0 to ε_m. For a given value of ε, the number of p-dimensional holes of the Vietoris-Rips complex $\mathcal{R}_\varepsilon$ is determined as the dimension of the vector space given by the pth homology group $H_p(\mathcal{R}_\varepsilon)$, where coefficients are taken to be in some fixed field, typically $\mathbb{Z}_2$. The *pth Betti number* is given by$\beta_p\left(\mathcal{R}_\varepsilon\right) = \dim\left[H_p\left(\mathcal{R}_\varepsilon\right)\right]$. However, even knowing the Betti numbers at all values of ε, one has no information on whether or not the corresponding topological features persist as ε increases. Persistent homology remedies this defect by encoding not just the Betti numbers, but the *persistent Betti numbers*, given by $\beta_i^j = \operatorname{rank}\left(H_p(\mathcal{R}_{\varepsilon_i}) \to H_p(\mathcal{R}_{\varepsilon_j})\right)$ where $H_p(\mathcal{R}_{\varepsilon_i}) \to H_p(\mathcal{R}_{\varepsilon_j})$ is the linear map induced by the inclusions $\mathcal{R}_{\varepsilon_i} \subset \mathcal{R}_{\varepsilon_j}$. The image of this linear map is called a *persistent homology group*. The *persistence diagram* gives a complete summary of persistent homology as a collection of points $\{(b,d)\}$, where each $(\varepsilon_i, \varepsilon_j)$ represents a homology class that is born at ε_i and dies at ε_j. To be precise, the multiplicity of the point $(\varepsilon_i, \varepsilon_j)$ in the persistence diagram is given by $\mu_i^j = \beta_{i-1}^j - \beta_i^j + \beta_i^{j-1} - \beta_{i-1}^{j-1}$. For a point (b,d) in the persistence diagram, the quantity $d - b$ is called its *persistence*.

31.2.1 Death vectors and persistence landscapes

In order to facilitate statistical inference we wish to give a complete (i.e. invertible) unique (i.e. injective) encoding of the persistence diagram as a vector. For the Vietoris-Rips complex, since all vertices appear at filtration value 0, all of the points in the persistence diagram for homology in degree 0 have birth coordinate 0. Thus, all of the information is included in the death times (the times when connected components merge). As such, we encode the persistence diagram using the corresponding order statistic. We call this the *death vector*. See the left hand side of Figure 1. For more general persistence diagrams, such as for homology in degree 1 for the Vietoris-Rips complex (see the right hand figure in Figure 1), we use the persistence landscape (Bubenik, 2015[2]), which we now describe. For each point

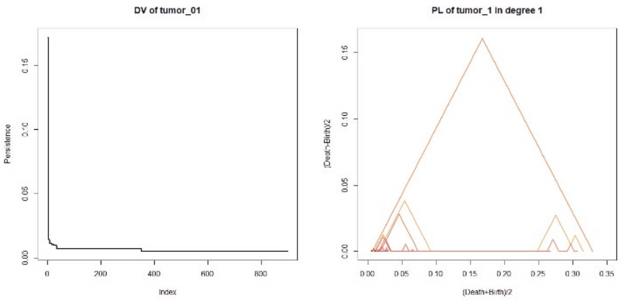


Fig. 31.1 Example of death vector(left) and persistence landscape(right).

(b, d) in the persistence diagram, consider the following function

$$f_{(b,d)}(t) = \begin{cases} t - b, & \text{if } b \le t < \frac{b+d}{2}, \\ d - t, & \text{if } \frac{b+d}{2} \le t < d, \\ 0, & \text{otherwise.} \end{cases}$$

Then for $k \geq 1$, the *kth persistence landscape function* of the persistence diagram $\mathcal{D}$ is given by

$$\lambda_k(t) = \text{kmaxChen}_{(b_i,d_i)\in\mathcal{D}}\, f_{(b_i,d_i)}(t),$$

where kmaxChen denotes the kth largest element. Looking from a *functional data perspective*, the *persistence landscape* consists of the sequence of functions $\{\lambda_1, \lambda_2, \lambda_3, \dots\}$. Note that by definition, for all $t \in \mathbb{R}$, $\lambda_1(t) \geq \lambda_2(t) \geq \lambda_3(t) \geq \dots$; that is, the persistence landscape is a decreasing sequence of functions, therefore one may apply data analysis techniques for a statistical inference.

31.2.2 Statistics with persistence landscapes

Bubenik and Kim (2007)[2] showed that persistence landscapes belong to a separable Banach space $L^p(\mathbb{N} \times \mathbb{R})$; and when $p \geq 2$ with finite first and second moments, asymptotic results are applicable. Let $X_1, \dots, X_n$ be iidr vectors and let $\Lambda_1, \dots, \Lambda_n$ be the corresponding persistence landscapes. Using the vector space structure, the sample mean landscape $\overline{\Lambda}_n$ is given by the pointwise mean. That is, $\overline{\Lambda}_n(\omega) = \overline{\lambda}_n$,where $\overline{\lambda}_n(k, t) = \frac{1}{n}\sum_{i=1}^{n} \lambda_i(k, t)$. On $L^p(\mathbb{N}\times\mathbb{R})$, when $p \geq 2$ and under the assumption of finite first and second moments, by delta method one can prove that for any continuous differentiable function f, the random variable $f(\lambda)$ also satisfies SLLN and CLT. One could therefore set an asymptotic $z - test$ to run statistical inference for persistence landscapes. Let $X_1, \dots, X_n$ be i.i.d.r, vectors from Q , and $X'_1, \dots, X'_{n'}$ be i.i.d.r, vectors from Q' with corresponding persistence landscapes Λ and Λ'. We set $Y = ||\Lambda||$, $Y' = ||\Lambda'||$. Let $\mu = EY$ and $\mu' = EY'$. Consider the null hypothesis $H_0 : \mu = \mu'$. For the corresponding real random variables $Y_n, Y'_{n'}$, we have $\sqrt{n}(\overline{Y_n} - \mu) \xrightarrow{d} N(0, \sigma_Y^2)$, and under H_0, similarly $\sqrt{n}(\overline{Y'}_{n'} - \mu) \xrightarrow{d} N(0, \sigma_{Y'}^2)$, where $\overline{Y}_n, \overline{Y'}_{n'}$ are the sample means and $\sigma_Y^2, \sigma_{Y'}^2$ are the corresponding variances. By Slutsky's theorem, if $S_n^2, S_{n'}^2$ are the corresponding sample variances, in case $\frac{n}{n+n'} \longrightarrow \theta \in (0, 1)$ as $n + n' \longrightarrow \infty$, we may use the statistic $Z = \frac{\overline{Y}_n - \overline{Y'}_{n'}}{\sqrt{S_n^2/n + S_{n'}^2/n'}}$ that converges weakly to standard normal random variable, as $n + n' \longrightarrow \infty$. From this nonparametric large sample z-like test, a p-value could be obtained. For small samples, we consider $X_1^*, \dots, X_n^*$ and $X_1'^*, \dots, X_{n'}^*$ be bootstrap resamples corresponding to samples from Q, Q' and corresponding norms of persistence landscapes resamples with repetition $Y_n^*, Y'_{n'}{}^*$, for both populations. Let $Y_n^*, Y'_{n'}{}^*$. By nonparametric bootstrap, we may obtain a two-sample test statistic, by taking for each resample from the two populations

$$Z^* = \frac{\overline{Y}_n^* - \overline{Y'}_{n'}^*}{\sqrt{S_n^{2*}/n + S_{n'}^{2*}/n'}}$$

Based on the nonparametric bootstrap distribution of Z^*, a p-value can be obtained.

31.2.2.1 Nonparametric χ^2 test for mean PL based difference

Alternately, one may consider a multivariate two sample test for mean PL's. Consider $\mathbf{Y} = (\mathbf{Y_1}, \ldots, \mathbf{Y_{n_1}})^T$, $\mathbf{Y_1} = (\|\Lambda_1^1| \ldots, \|\Lambda_1^p\|), \ldots, \mathbf{Y_{n_1}} = (\|\Lambda_{n_1}^1\|, \ldots, \|\Lambda_{n_1}^p\|)$; $\mathbf{Y'} = (\mathbf{Y'_1}, \ldots, \mathbf{Y'_{n_1}})^T$, $\mathbf{Y'_1} = (\|\Lambda'^1_1\|, \ldots, \|\Lambda'^p_1\|), \ldots, \mathbf{Y'_{n_2}} = (\|\Lambda'^1_{n_2}\|, \ldots, \|\Lambda'^p_{n_1}\|)$, where $p \ll n_1 + n_2$ represent the dimensionality of persistence landscape $\Lambda^1, \ldots, \Lambda^p$ that contains the information of persistence homology of the data in dimensions $1, \ldots, p$.

If the two groups have sample sizes $n_1 \neq n_2$, with total sample size $n = n_1 + n_2$, and $\lim_{n\to\infty} \frac{n_1}{n} = \theta \in (0,1)$, let $\overline{Y}_{n_1}, \overline{Y'}_{n_2}, S_{1,n_1}, S_{2,n_2}$ denote sample means and sample covariance matrices. Then

$$(\bar{Y}_{n_1} - \bar{Y'}_{n_2})^T [(\frac{1}{n_1} S_{1,n_1} + \frac{1}{n_2} S_{2,n_2})]^{-1} (\bar{Y}_{n_1} - \bar{Y'}_{n_2}) \to_d \chi_p^2 \text{as } n \to \infty.$$

For n large, we use the quantity on the left hand side to evaluate the p-value based on from the χ_p^2 tables. Similarly, if the sample sizes of the two PL groups are small, we use nonparametric bootstrap to form a statistic test: Let $Y_1^*, \ldots, Y_{n_1}^*$ and $Y_1'^*, \ldots, Y_{n_2}'^*$ be bootstrap resamples corresponding to p-dimensional persistence landscapes Λ,Λ'. If the samples sizes are small, by nonparametric bootstrap, we may use a test based on the χ^2 asymptotics. let $\overline{Y}^*_{n_1}, \overline{Y'}^*_{n_2}, S^*_{1,n_1}, S^*_{2,n_2}$ denote the bootstrap counterparts of sample means and sample covariances. Let $n = n_1 + n_2$ be the total sample size. Then, the boostrap distribution of the $T^{2,*}$ statistic

$$T^{2,*} = (\overline{Y}^*_{n_1} - \overline{Y'}^*_{n_2})^T [(\frac{1}{n_1} S^*_{1,n_1} + \frac{1}{n_2} S^*_{2,n_2})]^{-1} (\overline{Y}^*_{n_1} - \overline{Y'}^*_{n_2}), \tag{31.1}$$

can be used to derive a p-value for the $H_0 : \mu_Y = \mu_{Y'}$.

31.3 The GMB Images and their PL analysis

Our imaging data were collected of patients immediately before the pathological diagnosis, and from follow-up scans, where available. This data collection is from the NCI's Clinical Proteomic Tumor Analysis Consortium Glioblastoma Multiforme (CPTAC-GBM) cohort. Radiology and pathology images from CPTAC Phase 3 patients are being stored in DICOM format. Brain tumors are classified by the types of cells within the tumor. Each type of brain tumor grows and is treated in a different way. Glioblastoma is a type of Grade IV astrocytomas and the most common malignant (cancerous) adult brain tumor and one of the fastest-growing tumors of the central nervous system. Imaging tests, which include CT(computed tomography) scans and MRI(magnetic resonance imaging), are most common diagnostic tests for detecting brain tumors. In this study, CT scan images are used to study the GBM (see Figure 2). There are several levels of CT scans which detect different areas of normal anatomy of human brain. Two levels of CT scans are used here to form

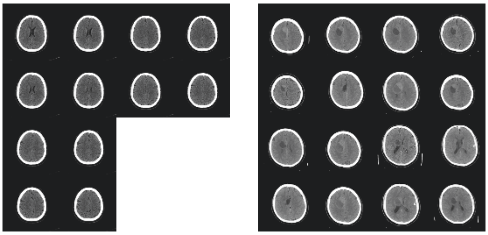


Fig. 31.2 Brain CT scans: normal (left) and tumor (right).

different groups for statistical analysis. The data used in the article contains total twenty-nine brain CT scan images. For Centrum semiovale level, eight images are from normal group and eleven images are from tumor group. For Lateral ventricles level, four images are from normal group and six images are from tumor group. Next, each of 29 images was extracted, in MATLAB with using Image Segmenter and edge map detection and then pairs of 2-dimensional points were selected and prepared for computations. Our computations were performed in MATLAB and Image Processing Toolbox Release 2016b, in R-3.5.2 with Jose Bouza's TDA-tools and in C++ with Ulrich Bauer's Ripser code. First image segmentation was performed in MATLAB's Image Segmenter with using Otsu's global threshold. Then, from segmented images, by using edge detection function, point clouds for each CT scan image were obtained. From these brain edges, point clouds of approximately 1000 points were sampled.

Next, the persistence diagrams for the Vietoris-Rips complexes of all the point cloud data sets were computed using Ripser. These persistence diagrams were then converted into vectors to facilitate statistical analysis. Specifically, the persistence diagrams for homology in degree 0 were converted into death vector and the persistence diagrams for homology in degree 1 were converted into persistence landscapes using Jose Bouza's tda-tools. Figure 4 shows the PLs for each sample from the normal group and GBM group. Next, we consider the average persistence landscapes for the normal group and tumor group and the differences between these averages. See Figure 5. The two groups highly differ in their topological type, as the p-value found was 0.007. Upon inspection, the difference between normal group and tumor group

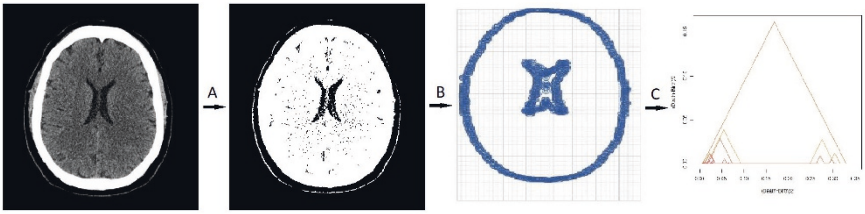


Fig. 31.3 Data Processing, A:Image Segmentation, B:Edge Detection, C:R-tda-tools

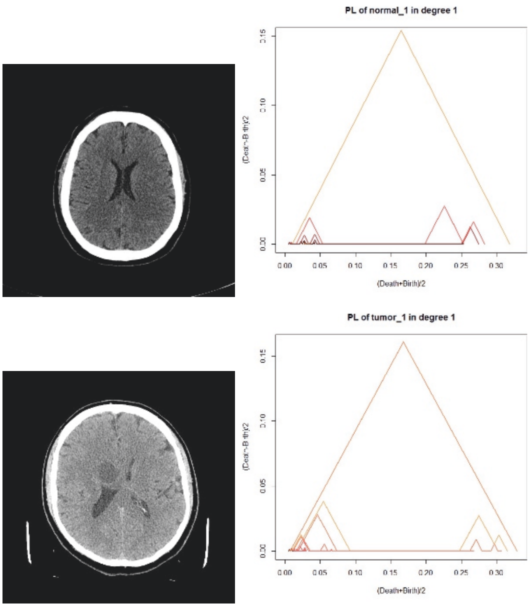


Fig. 31.4 CT scan of individual from normal group and corresponding PL (up), and CT scan of individual from GBM group and corresponding PL (down).

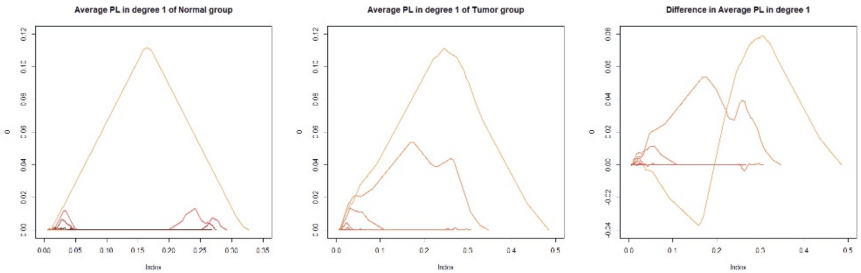


Fig. 31.5 Average PLs for normal group (left) and tumor group (middle) and their difference (right).

shows the two groups have difference topological features. A similar conclusion is reached, if the nonparametric T^2 like test in Section 2.2.1 is used instead. After computing the pooled sample covariance matrix, the distribution of the statistic (31.1) could be obtained, leading to a p-value smaller than .00001. In summary, CT scan based PLs, could be used to differentiate between the GBM group and the healthy group, using brain CT scan images.

References

[1] Bauer, Ulrich (2017). *Ripser: a lean C++ code for the computation of Vietoris–Rips persistence barcodes*. Software available at https://github.com/Ripser/ripser.

[2] Bubenik, P. and Kim, P.T. (2007). A Statistical Approach to Persistent Homology, *Homology, Homotopy and Applications*, 9(2), 337 – 362.

[3] Bubenik,P.(2015). Statistical Topological Data Analysis using Persistence Landscapes. *J. of Machine Learning Research*. **16**, 77–102.

[4] Bubenik, P. and Dlotko, P. (2017). A persistence landscapes toolbox for topological statistics.*A persistence landscapes toolbox for topological statistics*. **78**, 91 – 114.

[5] Bubenik, Peter; Carlsson, Gunnar; Kim, Peter T. and Luo, Zhi-Ming.(2010). Statistical topology via Morse theory persistence and nonparametric estimation. *Algebraic methods in statistics and probability II*, 75—92, *Contemp. Math.*, **516**, Amer. Math. Soc.

[6] Chen, Yen-Chi; Genovese, Christopher R.; Wasserman, Larry.(2017). Statistical inference using the Morse-Smale complex. *Electron. J. Stat.* **11**, 1390—1433

[7] Edelsbrunner, H. and Harer, J. (2008). Persistent Homology- a Survey. *Surveys on Discrete and Computational Geometry*. *Twenty Years Later*, eds. J.E. Goodman, J. Pach and R. Pollack, Contemporary Mathematics 453, Amer. Math. Soc., Providence, Rhode Island, 257 – 282.

[8] Edelsbrunner, Herbert; Harer, John L.(2010). *Computational topology. An introduction.* American Mathematical Society, Providence, RI. ISBN: 978-0-8218-4925-5

[9] H. Edelsbrunner, D. Letscher, A. Zomorodian (2002). Topological persistence and simplification. *Discrete & Computational Geometry* **28** (4), 511–533.

[10] L. Ellingson, F. H. Ruymgaart and V. Patrangenaru (2013). Nonparametric Estimation of Means on Hilbert Manifolds and Extrinsic Analysis of Mean Shapes of Contours. *Journal of Multivariate Analysis.* **122**, 317–333.

[11] Holland EC.(2000).Glioblastoma multiforme:the terminator. *Proc Natl Acad Sci USA. ,97(12):6242–6244.*

[12] Patrangenaru, V. and Ellingson, L. E. (2015). *Nonparametric Statistics on Manifolds and their Applications to Object Data Analysis.* CRC.

[13] V. Patrangenaru, P. Bubenik, R.Paige and D. Osborne (2018). Challenges in Topological Object Data Analysis . *Sankhya A: The Indian Journal of Statistics* https://doi.org/10.1007/s13171-018- 0137-7

[14] V. Patrangenaru, R. Paige, K. D. Yao, M. Qiu and D. Lester (2016). Projective Shape Analysis of Contours and Finite 3D Configurations from Digital Camera Image. *Statistical Papers.* **57**, 1017–1040.

Chapter 32
Distribution-free Pointwise Adjusted *P*-values for Functional Hypotheses

Meng Xu and Philip T. Reiss

Abstract Graphical tests assess whether a function of interest departs from an envelope of functions generated under a simulated null distribution. This approach originated in spatial statistics, but has recently gained some popularity in functional data analysis. Whereas such envelope tests examine deviation from a functional null distribution in an omnibus sense, in some applications we wish to do more: to obtain p-values at each point in the function domain, adjusted to control the family-wise error rate. Here we derive pointwise adjusted p-values based on envelope tests, and relate these to previous approaches for functional data under distributional assumptions. We then present two alternative distribution-free p-value adjustments that offer greater power. The methods are illustrated with an analysis of age-varying sex effects on cortical thickness in the human brain.

32.1 Introduction

In many functional data analysis (FDA) settings, one wishes to test either a null hypothesis

$$H_0 : f(s) = 0 \text{ for all } s \in \mathcal{S}, \tag{32.1}$$

for a function f defined on a domain $\mathcal{S}$, or alternatively a family of null hypotheses

$$\{H_0(s) : s \in \mathcal{S}\} \tag{32.2}$$

where for each s, $H_0(s)$ is the pointwise hypothesis $f(s) = 0$. For example, f may refer to

(i) a group difference $f(s) = g_1(s) - g_2(s)$, where g_1, g_2 denote mean functions in two subsets of a population, or

Meng Xu
Department of Statistics, University of Haifa, Haifa 31905, Israel, e-mail: mxu@campus.haifa.ac.il

Philip T. Reiss (✉)
Department of Statistics, University of Haifa, Haifa 31905, Israel, e-mail: reiss@stat.haifa.ac.il

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1_32

(ii) a coefficient function $f(s) = \beta(s)$ in a functional linear model.

Clearly the global hypothesis H_0 in (32.1) is just the intersection over all s of the pointwise hypotheses $H_0(s)$ in (32.2). The difference is that whereas (32.1) refers to a single test, for which a single p-value would be appropriate, the family (32.2) gives rise to a collection of p-values. The latter setup is appropriate when the values of $f(s)$ for different s carry distinct scientific meaning. For example, in Section 32.6 below we test for sex-related differences in the thickness of the human cerebral cortex as a function of age s. In this context, age-specific results may have implications for the study of brain development.

Previous work has tended to focus either on distribution-free tests of the global hypothesis (32.1) (see Section 32.3 below), or on multiplicity-adjusted parametric pointwise tests for the family (32.2). As we show in Section 32.4, it is straightforward to combine the advantages of both approaches—that is, to derive pointwise adjusted p-values without having to specify a null statistic distribution. In Section 32.5, we present two alternative pointwise p-value adjustments that offer improved power.

32.2 Setup

We let $T(s)$ ($s \in \mathcal{S}$) denote a functional test statistic for null hypothesis (32.1), and take as given a group of permutations of the data, along with the null hypothesis that the joint distribution of $T(s), s \in \mathcal{S}$, is invariant to such permutations. This hypothesis may be stronger than (32.1), but for the sake of a brief and general presentation, we ignore that distinction here. Let T_0 be the test statistic function computed with the real data, and $T_1, \ldots, T_{M-1}$ be test statistic functions that are computed with randomly permuted data sets and thus constitute a simulated null distribution. We consider $T_0(s), \ldots, T_{M-1}(s)$ only for $s \in \mathcal{G}$, for a finite set $\mathcal{G} \subset \mathcal{S}$ (e.g., a grid of points spanning $\mathcal{S}$, if the latter is a subinterval of the real line). We assume $\mathcal{G}$ to be an adequate approximation to $\mathcal{S}$, in the sense that the difference between a minimum over $\mathcal{G}$ versus over $\mathcal{S}$ is negligible (see [2] for a relevant treatment of grid approximations in functional hypothesis testing). We further assume that there are no pointwise ties, i.e., ties among $T_0(s), \ldots, T_{M-1}(s)$ for a given $s \in \mathcal{G}$.

32.3 Envelope Tests

Hypotheses regarding spatial point patterns are commonly tested by functions $T(s)$ of interpoint distance s, such as the K function of [15]. Such functions typically have unknown null distributions, and hence are most readily tested via Monte Carlo methods. This is the motivation for graphical or envelope tests [15, 3, 1], which have recently been formalized, extended, and applied to functional data [9, 8].

The global envelope test (GET) of [9] is based on the ranks $R_m^*(s)$ of $T_m(s)$ among $T_0(s), \ldots, T_{M-1}(s)$ for $s \in \mathcal{G}$. Here rank is defined in such a way that low rank indicates maximal inconsistency with the null hypothesis. Thus, depending on the test, $R_m^*(s)$ may be rank be from smallest to largest, rank from largest to smallest, or for a two-sided test, the smaller of the two. The minimum rank attained by T_m,

$R_m = \min_{s\in\mathcal{G}} R_m^*(s)$, is a functional depth [7], which we may call the min-rank depth. The GET p-value is then defined as

$$p_+ = \frac{\sum_{m=1}^{M-1} \mathbb{I}(R_m \leq R_0) + 1}{M}. \tag{32.3}$$

This p-value has a graphical interpretation in terms of envelopes, which we define here in a manner that is consistent with [9], but that relates to p-values rather than a specified level α. For $j \geq 1$, let $\kappa_j = \sum_{m=0}^{M-1} \mathbb{I}(R_m \leq j)$, and let E^{κ_j} be the envelope defined by the set of $M - \kappa_j$ curves $\{T_m : R_m > j\}$, that is, the range from $\underline{T}^{\kappa_j}(s) = \min_{m:R_m>j} T_m(s)$ to $\bar{T}^{\kappa_j}(s) = \max_{m:R_m>j} T_m(s)$ for each s. We say that T_0 exits this envelope at s if $T_0(s) \notin [\underline{T}^{\kappa_j}(s), \bar{T}^{\kappa_j}(s)]$. Arguing as in [9], one can show that $p_+ \leq \kappa_j/M$ if and only if T_0 exits E^{κ_j} at some s.

32.4 Adjusted p-values

Turning from the single hypothesis (32.1) to the family (32.2) of pointwise hypotheses, the naïve or raw permutation-based p-values are

$$p(s) = R_0^*(s)/M \tag{32.4}$$

for each s. These p-values, however, require adjustment for multiplicity [18] in order to control the overall type-I error rate, usually taken as the family-wise error rate (FWER). Strictly speaking, since the GET is a single test as opposed to a multiple testing procedure, adjusted p-values with respect to the GET are undefined. But it is natural to define the GET-adjusted p-value at s, in the notation of Section 32.3, as the smallest value κ_j/M such that T_0 exits the envelope E^{κ_j} at s. It can be shown that an equivalent definition is

$$\tilde{p}(s) = \frac{\sum_{m=1}^{M-1} \mathbb{I}[R_m \leq R_0^*(s)] + 1}{M}; \tag{32.5}$$

and that, as we would expect, the adjusted p-values $\tilde{p}(s)$ control the FWER.

The adjusted p-value (32.5) is not really new. The `fda` package [12] for R [11] offers permutation t- and F-tests for settings (i) and (ii), respectively, of the Introduction (and similar permutation F-tests are described by [14]). These tests yield pointwise adjusted p-values that are related to (32.5), but there are two differences. First, in the terminology of [4], the `fda` package offers *max T* adjusted p-values, whereas (32.5) is more akin to *min P* adjusted p-values, which are more appropriate when one cannot assume the null distribution of $T(s)$ to be identical across s. Second, [12] adopt a different permutation p-value convention in which the numerator and denominator are reduced by 1, leading to the zero p-value problem criticized by [10].

32.5 More Powerful p-value Adjustments

We describe next two alternative adjusted p-values that are bounded above by (32.5) and thus offer greater power.

32.5.1 Step-down Adjustment

In the language of multiple testing, the adjusted p-values (32.5) are of *single-step* type, suggesting that an analogous *step-down* procedure [17, 4, 16] would be more powerful. Define $S_i = \{s \in \mathcal{G} : R_0^*(s) \geq i\}$ for $i = 1, 2, \ldots$, and $R_{m;U} = \min_{s \in U} R_m^*(s)$ for $m \in \{0, \ldots, M-1\}$ and $U \subset \mathcal{G}$. We can then define the step-down adjusted p-value at s as

$$\tilde{p}^{\text{stepdown}}(s) = \max_{i \in \{1,\ldots,R_0^*(s)\}} \frac{\sum_{m=1}^{M-1} \mathbb{I}(R_{m;S_i} \leq i) + 1}{M}. \tag{32.6}$$

This expression is readily shown to be less than or equal to $\tilde{p}(s)$ in (32.5). Thus the step-down adjusted p-values offer greater power than their single-step counterparts, but they can be shown to retain control of the FWER.

32.5.2 Extreme Rank Length Adjustment

The min-rank depth R_m of Section 32.3 tends to be strongly affected by ties. In particular, typically $\kappa_1 > 1$ of the M functions attain rank 1 at some point and thus have $R_m = 1$, with the result that κ_1/M is the smallest attainable value of either p_+ or $\tilde{p}(s)$. An alternative functional depth, the *extreme rank length* (ERL), largely eliminates ties and thus leads to a more powerful variant of the GET. A formal definition of ERL appears in [9], but the basic idea is to break the tie among curves with the same min-rank depth R_m by ordering from longest to shortest extent of the region over which that minimum rank is attained. For example, four curves in Fig. 32.1 attain pointwise rank 1 (from the top) somewhere in the domain and thus all have $R_m = 1$; the ERL depths R_m^{ERL} =1-4, indicated in the figure, are based on the widths of these curves' regions of attaining rank 1.

An ERL envelope $E^{\kappa_j;\text{ERL}}$ [8] can be defined as in Section 32.3, but in terms of R_m^{ERL} rather than R_m. We can then proceed as in Section 32.4, and define $\tilde{p}^{\text{ERL}}(s)$, the ERL-adjusted p-value at s, as κ_j/M for the smallest κ_j such that $T_0(s)$ lies outside $E^{\kappa_j;\text{ERL}}$. This adjusted p-value is bounded above by (32.5), and hence offers improved power. However, unlike most p-value adjustments, the ERL adjustment is not order-preserving, in the sense that $p(s_1) > p(s_2)$ does not guarantee that $\tilde{p}^{\text{ERL}}(s_1) \geq \tilde{p}^{\text{ERL}}(s_2)$. An counterexample, that is, a pair of points s_1, s_2 for which $p(s_1) > p(s_2)$ but $\tilde{p}^{\text{ERL}}(s_1) < \tilde{p}^{\text{ERL}}(s_2)$, appears in Fig. 32.1. Some might argue that this non-order-preserving behavior vitiates the use of ERL-adjusted p-values altogether.

32.6 Application: Age-varying Sex Difference in Cortical Thickness

We consider cortical thickness (CT) measurements from a longitudinal magnetic resonance imaging study at the US National Institute of Mental Health, which were previously analyzed by [13]. Specifically, we examine CT in the right superior temporal gyrus in 131 males with a total of 355 observations, and 114 females with 300 observations, over the age range from 5–25 years (displayed in the left panel of Fig. 32.2). Viewing the observations as sparse functional data, we fit the model $y_i(s) = \beta_0(s) + \tau_i\beta_1(s) + \varepsilon_i(s)$, in which $y_i(s)$ is the ith participant's CT at age s; $\tau_i = 0, 1$ if this participant is male or female, respectively; and $\varepsilon_i(s)$ denotes error. We focus on testing whether the age-varying sex effect $\beta_1(s)$ (female minus male) equals zero; see the right panel of Fig. 32.2 for an estimate of this coefficient function, along with pointwise 95% confidence intervals.

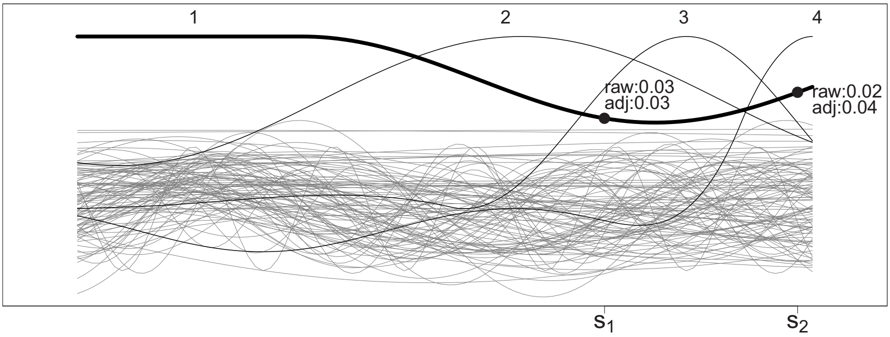


Fig. 32.1 An illustration of one-sided (higher = more extreme) ERL depths, and associated pointwise adjusted p-values. Here $M = 100$ and the numerals 1–4 denote ERL depths for the four curves with $R_m = 1$; the thickest curve represents the real data, so that $R_0^{\text{ERL}} = 1$. The raw p-values (32.4) satisfy $p(s_1) > p(s_2)$, but ERL adjustment reverses the order, i.e., $\tilde{p}^{\text{ERL}}(s_1) < \tilde{p}^{\text{ERL}}(s_2)$.

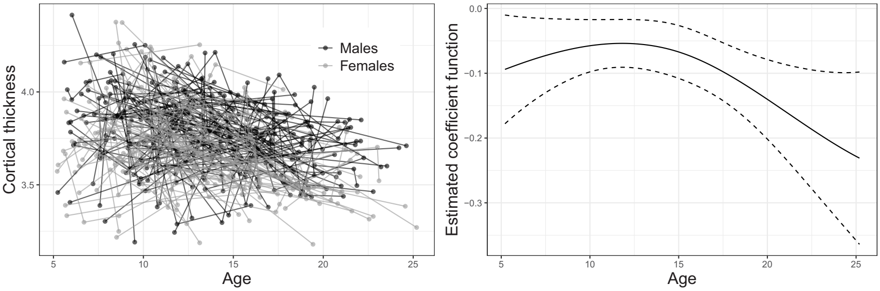


Fig. 32.2 Left: Cortical thickness in the right superior temporal gyrus for the NIMH sample. Right: Coefficient function estimate $\hat{\beta}_1(s)$ representing sex effect (female minus male), along with approximate pointwise 95% confidence interval.

The model was fitted by the pffr function [6], part of the R package refund [5], with both the real data and $M - 1 = 3999$ data sets with the sex labels permuted. The upper panel of Fig. 32.3 displays standardized coefficient functions $\hat{\beta}_1(s)/\ \widehat{\text{SE}}\ [\hat{\beta}_1(s)]$ for the real and permuted data sets, along with a two-sided envelope for testing at the 5% level. The GET p-value (32.3) based on min-rank depth is $p_+ = .003$; if we instead use the ERL depth, the GET p-value falls to .00025 (= $1/M$). But to quantify the evidence of a sex effect in an age-specific manner, we require pointwise p-values.

The lower panel of Fig. 32.3 shows the pointwise adjusted p-values $\tilde{p}(s)$ (32.5), along with the step-down and ERL-based adjusted p-values of Section 32.5, for an evenly spaced grid of 100 ages. Judging from the values of $\tilde{p}(s)$, there is only weak evidence of a CT difference between girls and boys up to age 9. The step-down p-values in this age range, on the other hand, are markedly lower and consistently below the conventional .05 level. The ERL-adjusted p-values are closer to $\tilde{p}(s)$ in this lower age range but, somewhat less visibly, are the lowest of the three p-values for age 16 and higher. Thus neither one of the two adjustments of Section 32.5 consistently dominates the other.

It must be acknowledged that the right superior temporal gyrus was specifically selected for the purpose of illustrating differences that may arise among the p-value adjustments. Comparable analyses for most other brain regions would have yielded less prominent differences.

32.7 Discussion

Expression (32.5) defines distribution-free pointwise adjusted p-values with respect to the global envelope test of [9]. A pointwise p-value approach such as this, which is

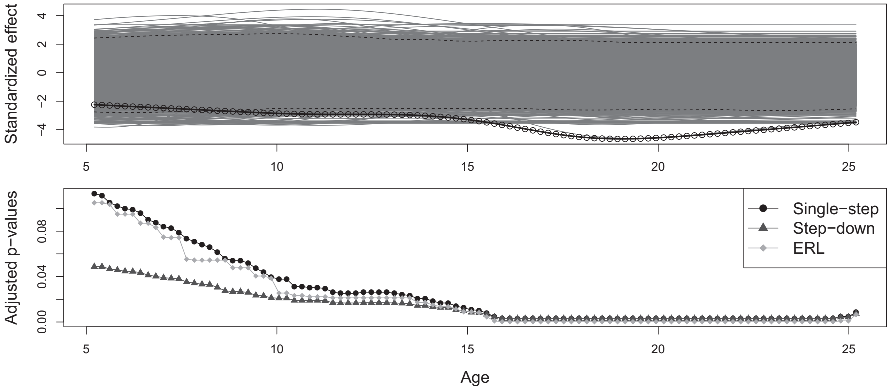


Fig. 32.3 Above: Standardized coefficient functions $\hat{\beta}_1(s)/\ \widehat{\text{SE}}\ [\hat{\beta}_1(s)]$ for the real data (black curve and circles) and for 3999 permuted data sets (grey curves), adapted from the R package GET [9]. Dashed lines indicate envelope for testing at the 5% level. Below: Pointwise adjusted p-values $\tilde{p}(s)$ (single-step), $\tilde{p}^{\text{stepdown}}(s)$ and $\tilde{p}^{\text{ERL}}(s)$.

agnostic with respect to the distribution of $T(s)$, is particularly valuable in analyses that go beyond pointwise t- or F-tests. For example, we are currently developing flexible pointwise tests for group differences in a measure of interest, based on estimating each group's density at each s, and then referring the distance between group-specific densities to a permutation distribution for each s; this distribution has no known analytic form under the null hypothesis.

The step-down and ERL-based adjusted p-values of Section 32.5 offer more powerful alternatives to (32.5), but some might question the suitability of the ERL adjustment since it is not order-preserving in general. The cortical thickness analysis of Section 32.6 illustrates the power gains that the step-down and ERL adjustments may provide in some applications. Simulation studies will further elucidate the relative performance of alternative p-value adjustments in FDA settings.

Acknowledgements This work was supported by Israel Science Foundation grant 1777/16. We thank Aaron Alexander-Bloch, Jay Giedd and Armin Raznahan for providing the cortical thickness data, and for advice on processing these data.

References

[1] Baddeley, A., Diggle, P.J., Hardegen, A., Lawrence, T., Milne, R.K., Nair, G.: On tests of spatial pattern based on simulation envelopes. Ecological Monographs **84**(3), 477–489 (2014)

[2] Cox, D.D., Lee, J.S.: Pointwise testing with functional data using the Westfall–Young randomization method. Biometrika **95**(3), 621–634 (2008)

[3] Davison, A.C., Hinkley, D.V.: Bootstrap Methods and Their Application. Cambridge University Press (1997)

[4] Ge, Y., Dudoit, S., Speed, T.P.: Resampling-based multiple testing for microarray data analysis (with discussion). TEST **12**(1), 1–77 (2003)

[5] Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Gellar, J., Harezlak, J., McLean, M.W., Swihart, B., Xiao, L., Crainiceanu, C., Reiss, P.T.: refund: Regression with Functional Data. R package version 0.1-17 (2018) https://CRAN.R-project.org/package=refund

[6] Ivanescu, A.E., Staicu, A.M., Scheipl, F., Greven, S.: Penalized function-on-function regression. Computational Statistics **30**(2), 539–568 (2015)

[7] López-Pintado, S., Romo, J.: On the concept of depth for functional data. Journal of the American Statistical Association **104**, 718–734 (2009)

[8] Mrkvička, T., Myllymäki, M., Jilek, M., Hahn, U.: A one-way ANOVA test for functional data with graphical interpretation. arXiv:1612.03608 (2018)

[9] Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H., Hahn, U.: Global envelope tests for spatial processes. Journal of the Royal Statistical Society: Series B **79**(2), 381–404 (2017)

[10] Phipson, B., Smyth, G.K.: Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. Statistical Applications in Genetics and Molecular Biology **9**(1), 39 (2010)

[11] R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2019) https://www.R-project.org/

[12] Ramsay, J.O., Hooker, G., Graves, S.: Functional Data Analysis with R and MATLAB. Springer, New York (2009)

[13] Reiss, P.T.: Cross-sectional versus longitudinal designs for function estimation, with an application to cerebral cortex development. Statistics in Medicine **37**(11), 1895–1909 (2018)

[14] Reiss, P.T., Huang, L., Mennes, M.: Fast function-on-scalar regression with penalized basis expansions. International Journal of Biostatistics **6**(1), 28 (2010)

[15] Ripley, B.D.: Modelling spatial patterns. Journal of the Royal Statistical Society: Series B **39**(2), 172–192 (1977)

[16] Romano, J.P., Wolf, M.: Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. Statistics & Probability Letters **113**, 38–40 (2016)

[17] Westfall, P.H., Young, S.S.: Resampling-Based Multiple Testing: Examples and Methods for *P*-Value Adjustment. John Wiley & Sons, New York (1993)

[18] Wright, S.P.: Adjusted *p*-values for simultaneous inference. Biometrics **48**(4), 1005–1013 (1992)

Authors Index

G. Aneiros et al. (eds.), *Functional and High-Dimensional Statistics and Related Fields*,
Contributions to Statistics, https://doi.org/10.1007/978-3-030-47756-1