II Joint Workshop on Functional Data Analysis and Nonparametric Statistics

La Cristalera (Miraflores de la Sierra, Madrid)
10–13 September, 2024

Proceedings of II Joint Workshop on Functional Data Analysis and Nonparametric Statistics, a meeting of the Functional Data Analysis and Nonparametric Statistics Working Groups of the Spanish Society of Statistics and Operations Research (SEIO)

Seio
Sociedad de Estadística e Investigación Operativa

UAM
Universidad Autónoma de Madrid

Last updated: 2024-09-16

Committees

Coordinators of SEIO working groups

Jose Ameijeiras Alonso – Nonparametric Statistics Working Group
Antonio Elías Fernández – Functional Data Analysis Working Group
Andrea Meilán Vila – Nonparametric Statistics Working Group
Luis Alberto Rodríguez Ramírez – Functional Data Analysis Working Group

Scientific Committee

José Ramón Berrendero Díaz – Universidad Autónoma de Madrid
José Enrique Chacón Durán – Universidad de Extremadura
Rosa María Crujeiras Casais – Universidade de Santiago de Compostela
María Dolores Ruíz Medina – Universidad de Granada

Organizing Committee

Jose Ameijeiras Alonso – Universidade de Santiago de Compostela
Antonio Elías Fernández – Universidad de Málaga
Eduardo García-Portugués – Universidad Carlos III de Madrid
Yolanda Larriba González – Universidad de Valladolid
Andrea Meilán Vila – Universidad Carlos III de Madrid
Luis Alberto Rodríguez Ramírez – Universidad Autónoma de Madrid
José Luis Torrecilla Noguerales – Universidad Autónoma de Madrid

Editors of book of abstracts: Organizing Committee

Workshop website: https://iiwfdanp.webnode.page/

Sponsors


100
101
001
Seio
Sociedad
de Estadística
e Investigación
Operativa


UAM
Universidad Autónoma
de Madrid


FUAM
Fundación
Universidad Autónoma
de Madrid


uc3m
Instituto Flores
de Lemus

Contents

Foreword

We are honored to present the Proceedings of the II Joint Workshop on Functional Data Analysis and Nonparametric Statistics (JW-FDA-NP). This workshop, organized in collaboration with the Scientific and Organizing Committees, represents the second joint meeting of the Functional Data Analysis (FDA) and Nonparametric Statistics (NP) Working Groups of the Spanish Society of Statistics and Operations Research (SEIO). Established in 2009 and 2019 respectively, these working groups aim to advance research in their respective fields under the SEIO umbrella.

Following the success of the inaugural JW-FDA-NP, this second joint edition also continues the trajectory set by the FDA Working Group through previous workshops held in Castro Urdiales (2019) and Getafe (2017, 2015). Additionally, this workshop marks the first collaboration with the NP Working Group.

The research intersections and common members between the FDA and NP communities are substantial, and hence the idea of gathering both in a joint workshop naturally appeared in the two working groups. The purpose is to bring together researchers who have contributed to each or both areas to disseminate their research, interact, and explore intersections. The workshop design dedicates specific days for each area, mixes senior and junior researchers, and tries to allocate enough time for the speakers to delve into details in the expositions of their works. As the venue for this special occasion, we have selected the Oberwolfach-esque residence of La Cristalera in Miraflores de la Sierra. We hope that this location can contribute to the event's success and create a memorable experience for the participants.

We express our gratitude to the authors of the abstracts of these proceedings for their highly-valuable scientific contributions. We gratefully acknowledge the members of the Scientific Committee for their efforts in ensuring the scientific quality of the workshop. We also thank the members of the Organizing Committee for their work in making the workshop a reality.

The workshop was possible due to the funding of the SEIO to the FDA and NP Working Groups, with the collaboration of the Instituto Flores de Lemus from the Universidad Carlos III de Madrid.

We hope you enjoy your time at the workshop!

The coordinators of SEIO Working Groups on FDA and NP
Madrid, 31 August 2024

Schedule

Monday 9 September

- **9:00 –** Dinner

Tuesday 10 September (FDA day)

- **9:00 – 9:15** Registration
- **9.15 – 9:40** Opening ceremony
- **9.40 – 9:45** Short break
- **9:45 – 11:15** **Short Course FDA (Part I)** **FDA** *Chair: Antonio Elías*
  - Data depth and applications (**Sara López Pintado**)
- **11:15 – 11:45** Coffee break
- **11:45 – 13:15** **Short Course FDA (Part II)** **FDA** *Chair: Antonio Elías*
  - Data depth and applications (**Sara López Pintado**)
- **13:15 – 13:40** **Contributed session** **FDA I** *Chair: Luis Alberto Rodríguez*
  - Diffusions with exact likelihood inference on the torus (**Eduardo García Portugués**)
- **13:40 – 15:10** Lunch
- **15:10 – 16:00** **Invited session** **FDA I** *Chair: José Luis Torrecilla*
  - Functional relevance based on continuous Shapley value (**Pedro Delicado**)
- **16:00 – 16:30** Coffe break
- **16:30 – 17:45** **Contributed session** **FDA II** *Chair: Antonio Elías*
  - 16:30 – 16:55. Variable selection based on non-differentiability of dependence measures (**Esther Jerez López**)
  - 16:55 – 17:20. A new scalar-on-function generalized additive model for partially observed functional data (**Pavel Hernández Amaro**)
  - 17:20 – 17:45. Predicting Electricity Supply and Demand Curves with Functional Data Techniques (**Zehang Li**)
- **17:45 – 21:00** Free time
- **21:00** Dinner

Wednesday 11 September (FDA day)

- **9:00 – 9:50** **Invited session** **FDA II** *Chair: José Luis Torrecilla*
  - 9:00 – 9:50. Estimating the dimension: some asymptotic theory and some empirical results (**Antonio Cuevas**)
- **9:50 – 10:40** **Contributed session** **FDA III** *Chair: Luis Alberto Rodríguez*
  - 9:50 – 10:15. Functional history matching for tsunami warnings (**Nicolás Hernández**)
  - 10:15 – 10:40. Wavelet Spatial ICA for Wide Functional Data (**Marc Vidal**)
- **10:40 – 11:10** Coffee break
- **11:10 – 12:50** **Invited session** **FDA III** *Chair: Luis Alberto Rodríguez*
  - 11:10 – 12:00. The Epigraph and Hypograph Indexes: Useful Tools in FDA (**Rosa Lillo**)
  - 12:00 – 12:50. Analysing the Complexity of Functional Data (**Enea Bongiorno**)
- **12:50 – 13:40** **Contributed session** **FDA-NP I** *Chair: José Luis Torrecilla*
  - 12:50 – 13:15. A kernel-based significance test for covariates in nonparametric functional regression (**Daniel Diz-Castro**)
  - 13:15 – 13:40. Conditional dependence testing with a functional application (**Laura Freijeiro González**)
- **13:40 – 15:30** Lunch
- **15:30 – 21:00** Social activity
- **21:00** Special dinner

Thursday 12 September (NP day)

- **9:00 – 10:30** Short Course NP (Part I) **NP** *Chairs: Jose Ameijeiras-Alonso y Andrea Meilán-Vila*
  - Dependence modeling with copulas (**Thomas Nagler**)
- **10:30 – 11:00** Coffee break
- **11:00 – 12:30** Short Course NP (Part II) **NP** *Chairs: Jose Ameijeiras-Alonso y Andrea Meilán-Vila*
  - Dependence modeling with copulas (**Thomas Nagler**)
- **12:30 – 13:20** Contributed session **NP I** *Chair: José E. Chacón*
  - 12:30 – 12:55. On uniqueness of the set of k-means (**Luis Alberto Rodríguez**)
  - 12:55 – 13:20. Kernel-based specification test for the ergression function (**María Vidal-García**)
- **13:20 – 14:50** Lunch
- **14:50 – 15:40** Invited session **NP I** *Chair: Jose Ameijeiras-Alonso*
  - Finding antimodes in univariate and bivariate data (**Jochen Einbeck**)
- **15:40 – 16:10** Coffe break
- **16:10 – 17:25** Contributed session **NP II** *Chair: Jose Ameijeiras-Alonso*
  - 16:10 – 16:35. Mode-based estimation of the center of symmetry (**José E. Chacón**)
  - 16:35 – 17:00. Highest density region estimation for manifold data (**Diego Bolón**)
  - 17:00 – 17:25. On an m-points class of uniformity tests on the sphere (**Alberto Fernández de Marcos**)
- **17:25 – 21:00** Free time
- **21:00** Dinner

Friday 13 September (NP day)

- **9:00 – 10:40** **Invited session** **NP II** *Chair: José E. Chacón*
  - 9:00 – 9:50. Recent advances in optimal transport-based inference methods (**Eustasio del Barrio**)
  - 9:50 – 10:40. Estimation and inference of panel data models with a generalized factor structure results (**Juan Manuel Rodríguez Poo**)
- **10:40 – 11:10** Coffee break
- **11:10 – 12:50** **Invited session** **NP III** *Chair: Andrea Meilán-Vila*
  - 11:10 – 12:00. Structured nonparametric density and ageperiod-cohort models (**María Dolores Martínez Miranda**)
  - 12:00 – 12:50. Presmoothed cure rate estimation in a mixture cure model (**María Amalia Jácome**)
- **12:50 – 13:40** **Contributed session** **NP III** *Chair: Andrea Meilán-Vila*
  - 12:50 – 13:15. Testing the effect of covariates on the cure rate using distance correlation (**Blanca Estela Monroy-Castillo**)
  - 13:15 – 13:40. Evaluation of nonparametric machine learning algorithms. A case of study in French Covid-19 data (**Germán Silva Gómez**)
- **13:40 – 14:00** Closing ceremony
- **14:00 – 16:00** Lunch

Abstracts: Shorts courses

Data depth and applications

10th Sept
09:45-13:15
FDA course

Sara Lopez-Pintado[1]
[1]Northeastern University

Data depth was originally introduced for multivariate data as a powerful non-parametric tool for developing robust exploratory data analysis methods. It measures how representative/central an observation is within the distribution or sample and provides a way of ranking observations from center-outward. Notions of depth have been extended to functional data in the last few decades. For example, the modified band depth introduced in [1] satisfies desirable properties and has been extensively used in many applications. Based on depth-rankings several outlier detection methods, robust classification procedures and rank tests have been developed. Depth-based methods such as an envelope test for detecting and visualizing differences between groups of functions have been recently proposed [3]. We have also introduced and established the properties of the metric halfspace depth, which is an extension of the well-known Tukey's depth to object non-Euclidean data in general metric spaces [3]. These novel data depth methods have been applied to several biomedical studies. For example, comparison of biometric data in normal versus premature babies, analysis of brain images and connectivity matrices in depression and Alzheimer disease, or studies of contagion curves in infectious processes under different network structures.

Keywords: data depth; functional data analysis; non-parametric statistics; robust statistics.

References

[1] S. Lopez-Pintado, J. Romo. *On the concept of depth for functional data*. Journal of the American Statistical Association. **104** (2009) 718-734.

[2] S. Lopez-Pintado, K. Qian. *A depth-based global envelope test for comparing two groups of functions with applications to biomedical data*. Statistics in Medicine. **40** (2021) 1639-1652.

[3] X. Dai, S. Lopez-Pintado. *Tukey's depth for object data*. Journal of the American Statistical Association. **118** (2023) 1760-1772.

12th Sept
09:00-12:30
NP course

Dependence modeling with copulas

Thomas Nagler[1,2]

[1]LMU Munich; [2]Munich Center for Machine Learning (MCML)

Copulas are powerful statistical tools for modeling and analyzing the dependence structure between random variables. This short course will provide a gentle introduction to copula models and some related association measures. After covering the fundamentals, we discuss their nonparametric estimation through the empirical copula process and kernel density estimators. Time permitting, we will conclude with a short excursion to vine copulas, which offer a flexible and efficient approach to modeling complex dependencies among multiple variables. This course is designed to equip participants with both theoretical knowledge and practical skills in dependence modeling.

Keywords: Dependence; Association; Estimation; Kernel.

Abstracts: Invited speakers

Recent advances in optimal transport-based inference methods

13th Sept
09:00-09:50
NP session II

Eustasio del Barrio[1]

[1]IMUVa, Universidad de Valladolid

Optimal transport has proven to be an important tool to compare probability measures since it enables to define a metric over the set of distributions which conveys their geometric properties. One of the central objects in the theory of optimal transport is the optimal transport (OT) map: the unique monotone transformation pushing forward an absolutely continuous probability law onto any other given law. OT maps play an important role in some recent statistical applications, either as a tool for defining multivariate analogues of quantile functions, for correcting distributional shifts in classification problems or in statistical inference over the space of probability measures. While, from the point of view of minimax rates, estimation of OT maps is affected by the curse of dimensionality, efficient estimation is possible in some interesting setups. In this talk we will present the case of entropic optimal transport ([3]) and that of semidiscrete optimal transport ([1]). We will illustrate the applicability of the results in the estimation of Laguerre cells and in Hotelling's location model for equillibrium prices.

Keywords: Optimal transport; Laguerre tessellations; Hotelling's location model

Acknowledgements This talk is based on joint work with A. González-Sanz and J. M. Loubes and on research partially supported by grant PID2021-128314NB-I00 funded by MCIN/AEI/ 10.13039/501100011033/FEDER, UE.

References

[1] E. del Barrio, A. González-Sanz, J. M. Loubes, J. Niles-Weed. *An improved central limit theorem and fast convergence rates for entropic transportation costs.* SIAM Journal on Mathematics of Data Science. **5** (2023) 639-669.

[2] E. del Barrio, A. González-Sanz, J. M. Loubes. *Central limit theorems for semi-discrete Wasserstein distances.* Bernoulli, **30** (2024) 554-580.

11th Sept
12:00-12:50
FDA session III

Analysing the Complexity of Functional Data

Enea G. Bongiorno[1], Kwo Lik Lax Chan[2], Aldo Goia[1], Philippe Vieu[3]
[1]Università del Piemonte Orientale - Amedeo Avogadro; [2]University of Birmingham;
[3]Université Toulouse III - Paul Sabatier

Analyzing data in high-dimensional or infinite-dimensional spaces often requires the use of dimensionality reduction techniques. One of the key challenges is determining the number of components (or degree of freedom) to consider. In this study, we propose a nonparametric approach to address this issue, leveraging the concepts of small-ball probability and complexity. The method will be illustrated through examples and practical applications.

Keywords: nonparametric, Small-ball factorization, complexity function, dimensionality reduction

References

[1] E.G. Bongiorno, L. Chan, A. Goia *Detecting the complexity of a functional time series*, J.Nonparametr.Stat. (2023) 1–23.

[2] E.G. Bongiorno, A. Goia, P. Vieu. *Estimating the complexity index of functional data: some asymptotics*. Statistics and Probability Letters, **161** (2020).

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

11th Sept
09:00-09:50
FDA session II

Estimating the dimension: some asymptotic theory and some empirical results

Antonio Cuevas[1], Alejandro Cholaquidis[2], Beatriz Pateiro-López[3]
[1]Universidad Autónoma de Madrid;[2]Universidad de la República;[3]Universidad de Santiago de Compostela

The problem of estimating, from a random sample of points, the dimension $\dim(S)$ of a compact subset S of the Euclidean space is considered. The emphasis is put on consistency results (in the statistical sense) of convergence to the true dimension value when the sample size grows to infinity, as well as on convergence rates. Among the many available definitions of $\dim(S)$, we have focused (on the grounds of its statistical tractability) on the *Minkowski dimension* and on the, perhaps less popular, concept of *pointwise dimension*. We prove the statistical consistency of some estimators proposed earlier in the literature by introducing other instrumental estimators formulated in terms of the empirical volume function $V_n(r)$, defined as the Lebesgue measure of the set of points whose distance to the sample is at most r.

In addition to these theoretical results, the outputs of an empirical study will be also commented.

A major statistical motivation for these dimension studies is the so-called "Manifold Hypothesis", according to which many real data sets found in practical applications are in fact grouped around a structure (typically a manifold) with a dimension smaller than that of the ambient space.

Functional relevance based on continuous Shapley value

10th Sept
15:10-16:00
FDA session I

Pedro Delicado[1] and Cristian Pachón-García[1]

[1]Universitat Politècnica de Catalunya

The presence of machine learning models in many facets of our lives has multiplied in recent years. Often, improvements in predictive efficiency come at the cost of increasing their complexity, which is why they are referred to as "black boxes". The opacity of certain algorithms has led to a growing demand to understand how and why they make their decisions. In response, a whole literature has recently emerged ("Interpretable Machine Learning" or "eXplainable Artificial Intelligence") whose goal is to make automatic algorithms transparent and interpretable. Among the interpretability methods, special attention has been given to those that are model agnostic (can be applied to any predictive model) and global (they measure the relevance/importance of each variable over the entire dataset). Multicolinearity among predictors makes it difficult to assign individual relevance measures to them. To overcome this problem, Lipovesky and Conklin (2001) proposed to adapt Shapley value imputation (a concept from cooperative games theory) to measure regressors' importance.

Consider a scalar-on-function regression problem, where the goal is to predict a scalar response from a functional predictor. Several predictive models have been proposed in the Functional Data Analysis literature: linear and nonlinear models, parametric and nonparametric, among others. In addition, other proposals have come from the machine learning literature: Support Vector Machine regression can be adapted to functional data, and versions of Neural Networks for functional data have recently appeared (Heinrichs et al., 2023). The above scalar-on-function predictive methods are generally difficult to interpret because it is hard to identify which features of the functional predictors are more important in computing the predicted values. In this work, we extend relevance measures based on the Shapley value from multivariate to functional predictors by adapting concepts from the continuous games literature. Our proposals are illustrated by a simulation study and several real data applications.

Keywords: Interpretable machine learning; scalar-on-function regression.

References

[1] S. Lipovetsky, M. Conklin . *Analysis of regression in game theory approach.* Applied Stochastic Models in Business and Industry **17** (2001) 319–330.

[2] F. Heinrichs, F., M. Heim, and C. Weber (2023). *Functional neural networks: Shift invariant models for functional data with applications to eeg classification.* In Proceedings of the 40 th International Conference on Machine Learning, Honolulu, Hawaii, USA.

12th Sept
14:50-15:40
NP session I

Finding antimodes in univariate and bivariate data

Jochen Einbeck[1]

[1]Durham University

For the estimation of density modes from real data, one of the available methods is the mean shift procedure, which essentially consititutes an iterative computation of a sequence of localized means. Beside the modes, that is, the maxima of a density function, data sets may also feature "antimodes", that is, local minima of the density function, or in other words, points "where the likelihood of observing data increases in any direction you move away from it". Methods to detect antimodes are less well developed in the statistical literature. In this talk, we will show how an inverse version of the mean shift procedure can be used to identify antimodes [1], and we will apply this methodology on univariate and bivariate data sets. A notable difficulty with this approach is that, unlike for the mean shift, convergence of the iterative procedure cannot be established. In fact, it can easily be demonstrated to be violated if the antimode is located in an area of extreme low density. However, this problem can be solved due to a simple algorithmic adaptation, as will be demonstrated in this presentation.

Keywords: modes; antimodes; kernel density estimation.

References

[1] J. Ameijeiras-Alonso, J. Einbeck. *A fresh look at mean-shift based modal clustering.* Advances in Data Analysis and Classification (2023), https://doi.org/10.1007/s11634-023-00575-1

Presmoothed cure rate estimation in a mixture cure model

13th Sept
12:00-12:50
NP session III

M. Amalia Jácome[1], Samuel Saavedra[1] and Ana López-Cheda[1]

[1]Universidade da Coruña (SPAIN)

Cure models are a particular case of models in survival analysis designed to accommodate the possibility that some individuals will never experience the event of interest. One of the main goals in cure models is to estimate the cure rate, that is, the probability that an individual belongs to the cured component of the population.

The nonparametric estimator for the cure rate is the Kaplan-Meier (KM) estimator, or the generalized KM estimator with covariates, evaluated at the largest uncensored time [3, 2]. Presmoothing has been shown to improve KM estimator [1], by replacing the indicators of no censoring with some preliminary nonparametric estimator of the conditional probability of uncensoring. The beneficial effect of presmoothing will be explored in the nonparametric estimator of the cure rate, and different bandwidth selectors will be compared. The resulting methods will be applied to a dataset related to breast cancer patients treated with potentially cardiotoxic oncologic drugs from the University Hospital of A Coruña (CHUAC) in 2007 - 2021. The potential impact on the cure rate of clinical covariates related to the aorta, the ventricles and the atria and image covariates such as the doppler echocardiogram will be analyzed.

Keywords: Bandwidth; Booststrap; Cross-validation; Kernel.

Acknowledgements MICINN Grant PID2020-113578RB-I00, Xunta de Galicia ED431C-2020-14, GAIN PCBAS MCSICAR project, and CITIC as a member of the CIGUS Network and co-financed by the EU ED431G 2023/01.

References

[1] R. Cao, M.A. Jácome. *Presmoothed kernel density estimation for censored data.* J. Nonpar. Stat. **16** (2004) 289-309.

[2] A. López-Cheda, R. Cao, M.A. Jácome, I. Van Keilegom. *Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models.* Comput. Stat. Data Anal. **105** (2017) 144-165.

[3] J. Xu, J. Y. Peng. *Nonparametric cure rate estimation with covariates.* Can. J. Stat. **42** (2014) 1-17.

11th Sept
11:10-12:00
FDA session III

The Epigraph and Hypograph Indexes: Useful Tools in FDA

Rosa E. Lillo[1], Belén Pulido[2] and Alba Franco-Pereira[3]

[1]uc3m-Santander Big Data Institute. Department of Statistics, uc3m; [2]uc3m-Santander Big Data Institute, uc3m;[3] Department of Statistics, ucm

As it is known, there is no natural order in functional data, which complicates the extension of some classic data analysis techniques from the real line to the functional domain. To address this deficiency, several ordering options have appeared in the literature, the most well-known being functional depth, with many variants aiming to order data from the center outward. An alternative that emulates the natural ascending or descending order on the real line is the order induced by the epigraph and hypograph indexes. This talk will describe the power of these indexes as tools for performing various activities related to functional data, such as rank tests, clustering, and outlier detection. The proposed methodology will be illustrated with simulated data and real examples.

Keywords: Epigraph and hypograph indexes; clustering FDA, outlier detection; multivariate FDA.

Acknowledgements

This research is part of the I+D+i projects PDC2022-133359, PID2022-137243OB-I00 and TED2021-131264B-100 funded by MCIN/AEI/10.13039/501100011033 and European Union NextGenerationEU/PRTR. This initiative has also been partially carried out within the framework of the Recovery, Transformation and Resilience Plan funds, financed by the European Union (Next Generation) through the grant ANTICIPA.

References

[1] B. Martín-Barragán, R. E. Lillo, J. Romo. *Functional Boxplots based on epigraphs and hypographs*, Journal of Applied Statistics. **43(6)** (2016) 1088–1103.

[2] A. Franco-Pereira, R. E. Lillo. *Rank tests for functional data based on the epigraph, the hypograph and associated graphical representations*, Data Analysis and Classification. **14(3)** (2020) 1088–1103.

[3] B. Pulido, A. Franco-Pereira, R. E. Lillo. *A fast epigraph and hypograph-based approach for clustering functional data*, Statistics and Computing. **33**, article number 36 (2023) 1088–1103.

Structured nonparametric density and age-period-cohort models

13th Sept
11:10-12:00
NP session III

María Luz Gámiz[1], Enno Mammen[2], *María Dolores Martínez-Miranda*[1] and Jens Perch Nielsen[3]

[1]University of Granada (Spain); [2]Heidelberg University (Germany); [3] Bayes Business School, City, University of London (UK)

Age-period-cohort (APC) models are important structures used to model (for example) demographic, economic, medical, behavioral and scientific output developments over time. From a nonparametric perspective, APC models can be seen as structured nonparametric density models ([2], [1], [3]), where the classical approach in the literature involves histogram-type estimators. In this talk we will describe this connection and present a generalisation of APC models allowing for time acceleration in the age direction. We call the new class of models AAPC for accelerated age-period-cohort models. This new AAPC class of models comes with simple solutions to identification of the past, permissible extrapolations for the future and statistical validation: three current research challenges even in the well known APC models. The new methodology is illustrated via the important case of understanding future fertility.

Keywords: Structured density; APC model; In-sample forecasting; Validation.

Acknowledgements This work has been partially supported by ERDF / Spanish Ministry of Science and Innovation - State Research Agency, through the grant PID2020-116587GB-I00.

References

[1] Lee, Y. K., Mammen, E., Nielsen, J. P., and Park, B. P. *Asymptotics for In-Sample Density Forecasting.* Annals of Statistics **43** (2015) 620–651.

[2] Mammen, E., Martínez-Miranda, M. D. and Nielsen, J. *In-sample forecasting applied to reserving and mesothelioma mortality.* Insur. Math. Econ. **61** (2015) 76–86.

[3] Mammen, E., Martínez-Miranda, M. D., Nielsen, J.P. and Vogt, M. *Calendar effect and in-sample forecasting.* Insur. Math. Econ. **96** (2021) 31–52.

13th Sept
09:50-10:40
NP session II

Estimation and inference of panel data models with a generalized factor structure

Juan M. Rodríguez-Poo[1], Alexandra Soberón[1], Stefan Sperlich[2]
[1]Universidad de Cantabria; [2]Université de Geneve

This paper introduces a novel panel data model with a fairly general structure for the unobserved common factors which also encompasses both, traditional additive and interactive fixed effects. Under rather weak assumptions, we obtain consistent estimators that are asymptotically normal at rate root-NT for the parameters of interest, while optimal nonparametric estimators are obtained for the unspecified part. Furthermore, the statistical properties of the resulting estimators are robust to misspecification of the relationship between common factors and factor loadings. We provide a nonparametric specification test for the crucial modeling assumption. It relies on combining the methodology of conditional moment tests and nonparametric estimation techniques. Using degenerate and nondegenerate theories of U-statistics we can show its convergence and asymptotically distribution under the null, and that it diverges under the alternative at a rate arbitrarily close to $\sqrt{NT}$. Finite sample inference is based on bootstrap. Simulations reveal an excellent performance of our methods. They are used to study the effect of the European Union Emissions Trading System on CO2 emission and economy of some countries of the European Union.

Abstracts: Contributed speakers

Highest density region estimation for manifold data

12th Sept
16:35-17:00
NP session II

Diego Bolón [1], Rosa M. Crujeiras [2] and Alberto Rodríguez-Casal[2]

[1]Université Libre de Bruxelles, Brussels, Belgium;[2]Galician Center for Mathematical Research and Technology, CITMAga, Universidade de Santiago de Compostela, Santiago de Compostela, Spain.

Highest density regions (HDRs) are the sets where the density function of the data exceeds a given (and usually high) threshold. Estimating the HDRs of a population from a data sample has many practical applications, including data clustering [3] and seismic data analysis [2]. While HDR estimation for Euclidean data has been well-studied, to the best our knowledge, [4] and [1] provide the only two proposals for HDR estimation for manifold data. This two contributions propose a plug-in estimator, which ignores the geometric structure of the problem. To address this issue, a new non-parametric HDR estimator for manifold data has been developed. This approach combines an underlying density estimator with some prior geometric information that is incorporated in the estimation method to simplify the final estimator. Specifically, the new proposal can be viewed as a generalization of the euclidean HDR estimation technique introduced by [5] to Riemannian manifolds. The consistency of the new estimator will be proven, and its convergence rate will be derived. Finally, the performance in practice of the new HDR estimator is illustrated with a real data example.

Keywords: non-parametric statistics; set estimation; highest density regions; manifold data.

References

[1] Cholaquidis, A., Fraiman, R., and Moreno, L. *Level set and density estimation on manifolds.* Journal of Multivariate Analysis. **189** (2022) 104925.

[2] Huo, X., and Lu, J.C. *A network flow approach in finding maximum likelihood estimate of high concentration regions.* Computational Statistics & Data Analysis, **46** (2004) 33–56.

[3] Rinaldo, A., and Wasserman, L. *Generalized density clustering.* The Annals of Statistics. **38** (2010) 2678–2722.

[4] Saavedra-Nieves, P., and Crujeiras, R.M. *Nonparametric estimation of direc- tional highest density regions.* Advances in Data Analysis and Classification. **16** (2021) 761–796.

[5] Walther, G. *Granulometric smoothing.* The Annals of Statistics. **25** (1997) 2273–2299.

12th Sept
16:10-16:35
NP session II

Mode-based estimation of the center of symmetry

José E. Chacón[1] and Javier Fernández-Serrano[2]
[1]Universidad de Extremadura; [2]Universidad Autónoma de Madrid

In the mean-median-mode triad of univariate centrality measures, the mode has been overlooked for estimating the center of symmetry in continuous and unimodal settings. This talk expands on the connection between kernel mode estimators and M-estimators for location, bridging the gap between the nonparametrics and robust statistics communities. The variance of modal estimators is studied in terms of a bandwidth parameter, establishing conditions for an optimal solution that outperforms the household sample mean. A purely nonparametric approach is adopted, modeling heavy-tailedness through regular variation. The results lead to an estimator proposal that includes a novel one-parameter family of kernels with compact support, offering extra robustness and efficiency. The effectiveness and versatility of the new method are demonstrated in a real-world case study and a thorough simulation study, comparing favorably to traditional alternatives.

Keywords: kernel mode estimator; center of symmetry; unimodality; redescending M-estimator; efficient nonparametric estimation.

Acknowledgements The research of the first author has been supported by the MICINN grant PID2021-124051NB-I00.

11th Sept
12:50-13:15
FDA-NP session I

A kernel-based significance test for covariates in nonparametric functional regression

Daniel Diz-Castro[1], Manuel Febrero-Bande[1,2] and Wenceslao González-Manteiga[1,2]
[1]Department of Statistics, Mathematical Analysis and Optimization. University of Santiago de Compostela, Spain; [2]Galician Centre for Mathematical Research and Technology (CITMAga), Santiago de Compostela, Spain.

The aim is to develop a test with bootstrap calibration to check the conditional mean independence between a response and a set of covariates, given another set of covariates which are assumed to be relevant in a nonparametric regression model. The test is derived from the "Kernel-based Conditional Mean Dependence" proposal in [1] and relies on nonparametric kernel estimation for the regression function under the null hypothesis. It is shown that the proposed test is consistent against unspecified alternatives under some mild regularity conditions and is able to detect local alternatives approaching the null hypothesis at $\sqrt{n}$-rate. Both the response and the set of covariates whose contribution to the regression model is being tested are allowed to take values in Hilbert spaces. The proposal is illustrated with a simulation study to evaluate its finite sample performance.

Keywords: significance test; conditional mean independence; functional data; bootstrap.

Acknowledgements The authors acknowledge support from project PID2020-116587GB-I00 funded by MCIN/AEI/10.13039/501100011033 and the Competitive Reference Groups 2021-2024 (ED431C 2021/24) from the Xunta de Galicia.

References

[1] T. Lai, Z. Zhang, Y. Wang. . *A kernel-based measure for conditional mean dependence*. Comput. Stat. Data Anal. **160** (2021), 107246.

On an m-points class of uniformity tests on the sphere

12th Sept
17:00-17:25
NP session II

Alberto Fernández-de-Marcos[1], Eduardo García-Portugués[1]

[1]Department of Statistics, Universidad Carlos III de Madrid

When testing uniformity on the hypersphere, the Sobolev class of tests is arguably one of the most comprehensive. The test statistics belonging to such class are V-statistics of order two whose structure can be interpreted as an integral of the quadratic difference between a given function evaluated in the sample and one, or as an expansion in spherical harmonics of a given bivariate kernel. Exploiting this second equivalence, we introduce a generalization of the Sobolev class of test statistics with kernels of order $m \geq 2$. These new tests can explore interactions between m points of the sample, detecting dependence structures elusive to Sobolev tests. The asymptotic null distribution for these m-tests is obtained and shown to be applicable in practice. The computational complexity of the new class is investigated. The m-tests are compared to the Sobolev tests under different alternatives, illustrating the advantage of the former in certain scenarios.

Keywords: Sobolev tests; Spherical data; Uniformity tests.

Acknowledgements The authors are supported by grant PID2021-124051NB-I00, funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe". The second author acknowledges support from "Convocatoria de la Universidad Carlos III de Madrid de Ayudas para la recualificación del sistema universitario español para 2021–2023", funded by Spain's Ministerio de Ciencia, Innovación y Universidades.

11th Sept
13:15-13:40
FDA-NP session I

Conditional dependence testing with a functional application

Laura Freijeiro-González[1,3], Wenceslao González-Manteiga[2,3] and Manuel Febrero-Bande[2,3]

[1]Universidad de Oviedo (UNIOVI); [2]CITMAga; Universidade de Santiago de Compostela (USC); [3] Grupo MODESTYA, Departamento de Estatística, Análise Matemática e Optimización

Conditional dependence is a distinct type of relationship compared to the usual concept of dependence. Given two random variables, X and Y, these are dependent conditioned to a third one, Z. Therefore, X and Y could be independent but have a conditional dependence relation given Z. We briefly review this implication and introduce a novel measure of conditional dependence: the conditional distance covariance correlation (CDC) of [1]. A discussion about its high computational cost and strategies to make its estimation possible in practice follows. In particular, we discuss the proper selection of bandwidth parameters for estimation and calibration in the vectorial framework of conditional dependence. Furthermore, we propose a new method for choosing automatic bandwidths for both and demonstrate the improvements of this selection through a simulation study. Finally, we provide some insights on how to apply these concepts in a concurrent functional regression model for conducting specification tests.

Keywords: bandwidth selection; concurrent functional regression model; conditional dependence; conditional distance correlation; specification tests.

Acknowledgements Authors acknowledge the support from the R&D project PID2020-116587GB-I00 granted by MICIU/AEI/10.13039/501100011033 and by the "ERDF A way of making Europe" and the Competitive Reference Groups 2021-2024 (ED431C 2021/24) from the Xunta de Galicia through the ERDF. We also acknowledge to the Centro de Supercomputación de Galicia (CESGA) for the computational resources provided.

References

[1] X. Wang, W. Pan, W. Hu, Y. Tian, H. Zhang. *Conditional distance correlation.* JASA. **110(512)** (2015) 1726–1734.

Diffusions with exact likelihood inference on the torus

10th Sept
13:15-13:40
FDA session I

Eduardo García-Portugués[1], Michael Sørensen[2]
[1]Universidad Carlos III de Madrid; [2]University of Copenhagen

We provide a class of stochastic differential equations on the torus with explicit transition probability densities, enabling exact likelihood inference. The presented diffusions are time-reversible and can be constructed for any pre-specified stationary distribution on the torus, including highly-multimodal mixtures. We give results on asymptotic likelihood theory allowing one-sample inference and tests of linear hypotheses for k groups of diffusions, including homogeneity. We show that exact and direct diffusion bridge simulation is possible too. The new family of diffusions is applied to (i) test several homogeneity hypotheses on the movement of ants and (ii) construct diffusion bridges between related proteins. If time allows, a class of circular jump processes with similar properties will be discussed.

Keywords: Diffusions; Likelihood; Protein structure.

Acknowledgements The first author acknowledges support from grant PID2021-124051NB-I00, funded by MCIN/AEI/10.13039/50110001103 and by "ERDF A way of making Europe". His research was also supported by "Convocatoria de la Universidad Carlos III de Madrid de Ayudas para la recualificación del sistema universitario español para 2021–2023", funded by Spain's Ministerio de Ciencia, Innovación y Universidades. Part of this work was developed during the Oberwolfach workshop "Statistics of Stochastic Differential Equations on Manifolds and Stratified Spaces"; both authors gratefully acknowledge the hospitality of the organizers and MFO.

Functional history matching for tsunami warnings

11th Sept
9:50-10:15
FDA session III

Ryuichi Kanai[1,2], *Nicolás Hernández*[3], Devaraj Gopinathan[4], Serge Guillas[1,2]
[1]Department of Statistical Science, University College London, London, United Kingdom; [2]Alan Turing Institute, London, United Kingdom;[3]Data Science, Statistics and Probability Centre, School of Mathematical Sciences, Queen Mary University of London, London, United Kingdom;[4]Advanced Research Computing Centre, University College London, London, United Kingdom

Tsunamis are catastrophic natural disasters that cause immense impacts on human lives and economies; however, a complete forecasting methodology has yet to be achieved. Modelling this phenomena and executing simulations to generate predictions have become one of the key task in the field. However, accurately determining the appropriate input values for simulation models beforehand presents significant challenges, particularly as simulations grow in complexity and sophistication. A method known as history matching offers a solution by utilising emulation to separate plausible input value regions from implausible regions. While history matching has been conventionally applied to discrete data, no equivalent methods previously existed for functional data, which underpin the natural phenomena of Tsunamis. To bridge this gap, we have devised a novel functional history matching technique tailored specifically for this type of data.

This innovative method integrates sophisticated techniques, including the Outer Product Emulator for functional emulation and Random Projection for dimensionality reduction. Furthermore, our approach facilitates the evaluation of derivative information, thus enhancing the precision in estimating appropriate model input values compared to traditional discrete history matching or functional principal component analysis, a highly esteemed method within functional data analysis.

To ascertain the effectiveness of this new approach, we conducted an analysis using time-series observational data for tsunami forecasts. The results of this analysis affirm that our method is exceedingly effective for natural phenomena that exhibit inherently continuous data structures.

Keywords: Functional data analysis, Random projection, Stochastic process, Model discrepancy, Functional emulator, History Matching, Tsunami simulation.

Acknowledgements The first author acknowledges support from grant PID2021-124051NB-I00, funded by MCIN/AEI/10.13039/50110001103 and by "ERDF A way of making Europe". His research was also supported by "Convocatoria de la Universidad Carlos III de Madrid de Ayudas para la recualificación del sistema universitario español para 2021–2023", funded by Spain's Ministerio de Ciencia, Innovación y Universidades. Part of this work was developed during the Oberwolfach workshop "Statistics of Stochastic Differential Equations on Manifolds and Stratified Spaces"; both authors gratefully acknowledge the hospitality of the organizers and MFO.

A new scalar-on-function generalized additive model for partially observed functional data

10th Sept
16:55-17:20
FDA session II

Pavel Hernández Amaro[1], Maria Durbán[1] and M. Carmen Aguilera-Morillo[2]

[1]Universidad Carlos III de Madrid; [2]Universitat Politècnica de València

Functional data analysis is one of the fastest growing fields in statistical analysis. Functional data is usually found as discrete and noisy observations of the true underlying function, measured over some continuum. In most cases, it is assumed that all functions are observed over the full extension of their domain. However, in many real data sets, each curve is observed in a subset of the domain, which may be different for each curve. This type of data is known as partially observed functional data.

In this work, we present a new methodology to fit a generalized scalar-on-function regression model for partially observed functional data. The proposed model considers each curve only within its observed subset of the domain; also a penalty is added to the estimation of the functional coefficient in order to control its smoothness through the smoothing parameter. Additionally, a basis representation of the functional data and the functional coefficient of the model is assumed. As a consequence, the functional model can be formulated as a mixed effect model, estimating directly all the model coefficients. This model has been extended to deal with two or more regressors with different dimensions, i.e., curves and surfaces.

Keywords: Partially observed functional data; scalar-on-function regression model; Basis representation.

Acknowledgements This work is supported by the grant PID2022-137243OB-I00. from the Spanish Ministry of Science, Innovation and Universities MCIN/AEI/10.13039/501100011033.

Variable selection based on non-differentiability of dependence measures

10th Sept
16:30-16:55
FDA session II

Esther Jerez López[1], José Ramón Berrendero Díaz[1] and José Luis Torrecilla Noguerales[1]

[1]Universidad Autónoma de Madrid

We consider the binary classification of Gaussian functional data and assume that the relevant information is contained in the values of the functions at a finite number of points (the so-called impact points). In this setting, we suggest a new variable selection method for identifying the impact points so that it is possible to replace each functional observation with a finite-dimensional vector without losing information.

Many existing variable selection methods are based on dependency measures between the response variable and the function values at each point. In particular, a direct relationship between impact points and the non-differentiability of the covariance function has been found recently in the regression setting (see [1]). We conjecture that this property can be extended to the classification problem and to other, more complex dependency measures capable of detecting non-linear relationships, such as the distance covariance [3]. Specifically, we have shown that the set of points where the distance covariance is non-differentiable includes the set of impact points. This result indicates that the identification of impact points is a problem closely related to the detection of points where the distance covariance is non-differentiable. Our method adapts the algorithm given in [1] to detect the non-differentiability points of the distance covariance function. The performance of the method is studied through simulations under several scenarios.

Keywords:Functional data analysis, functional classification, impact points, variable selection, distance covariance

Acknowledgements Special acknowledgements to the Universidad Autónoma de Madrid which is financially supporting the research with a FPI-UAM scholarship.

References

[1] D. Poß, D. Liebl, A. Kneip, H.Eisenbarth, T. D. Wager, L. F. Barrett. *Super-consistent estimation of points of impact in nonparametric regression with functional predictors*. Journal of the Royal Statistical Society Series B: Statistical Methodology **82(4)** (2020) 1115-1140.

[2] G. J. Székely, M. L. Rizzo, N. K. Bakirov. *Measuring and testing dependence by correlation of distances*. The Annals of Statistics, **35(6)** (2007) 2769-2794.

10th Sept
17:20-17:45
FDA session II

Predicting Electricity Supply and Demand Curves with Functional Data Techniques

Zehang Li[1], Andrés M. Alonso[2] and Lorenzo Pascual[3]

[1]Department of Statistics, Universidad Carlos III de Madrid; [2]Department of Statistics and IFL, Universidad Carlos III de Madrid;[3]IE Business School

In any liberalized energy market, forecasts of hourly electricity prices and accurate estimates of supply and demand curves are crucial. This ensures system stability and contributes to a more competitive and efficient market. Most articles applied univariate time series, machine learning, and deep learning models, predicting future prices based on their past and various explanatory variables [1, 2, 3]. Fewer studies predict considering the market-clearing mechanism behind the prices. [4] and [5] proposed two-step forecasting procedures by first estimating the supply and demand curves, and then intersecting these curves for hourly price prediction.

Our proposal improves the two-step approaches by combining functional PCA with time series modelling and functional regression analysis. In particular, we incorporate (1) curves in regular and seasonal lags as functional covariants considering strong temporal dependency, (2) meteorological information from all Spanish provinces as additional independent variables, (3) step basis functions to avoid impractical approximations and predictions, and (4) a monotonicity correction procedure adjusting a prediction with the closest curve from the training set. We conduct an exhaustive backtesting exercise for predicted supply and demand curves and the resulting hourly prices. In both cases, the out-of-sample predictive measures obtained for absolute errors are very promising.

Keywords: Electricity market; Supply curve; Demand curve; Functional regression analysis; Functional principal component analysis.

Acknowledgements The authors gratefully acknowledge financial support from the Spanish government through the Ministry of Science and Innovation projects PID2019-108311GB-I00 and PID2022-138114NB-I00.

References

[1] J. Nowotarski, R. Weron. *Recent advances in electricity price forecasting: A review of probabilistic forecasting*. Renewable and Sustainable Energy Reviews. **81** (2018) 1548–1568.

[2] T. Hong, P. Pinson, Y. Wang, R. Weron, D. Yang, H. Zareipour. *Energy Forecasting: A Review and Outlook*. IEEE Open Access Journal of Power and Energy. **7** (2020) 376-388.

[3] J. Lago, G. Marcjasz, B. De Schutter, R. Weron. *Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark*. Applied Energy. **293** (2021) 116983.

[4] F. Ziel, R. Steinert. *Electricity price forecasting using sale and purchase curves: The X-Model*. Energy Economics. **59** (2016) 435-454.

[5] M. Soloviova, T. Vargiolu. *Efficient representation of supply and demand curves on day-ahead electricity markets.* Journal of Energy Markets. **14** (2021) 99-126.

Testing the effect of covariates on the cure rate using distance correlation

13th Sept
12:50-13:15
NP session III

Blanca Monroy-Castillo [1], Ingrid Van Keilegom [2], Amalia Jácome[1], and Ricardo Cao [1]

[1]Universidade da Coruña (Spain); [2] KU Leuven (Belgium)

The concept of dependence among observations plays a central role in many fields. In survival analysis, measuring the relationship between the lifetime and a covariate is usually of major interest. One of the goals in cure models [3] is to test whether a covariate influences the cure rate.

Distance correlation [1] is a novel class of multivariate dependence coefficients with advantages over classical correlation coefficients: it is applicable to random vectors of arbitrary dimensions not necessarily equal, and it is zero if and only if the vectors are independent . Different estimators have been compared in [3].

In a standard survival model, [4] proposed an estimator for the distance covariance between covariates and survival times under right-censoring. But to the best of our knowledge, distance correlation has not been applied yet in the presence of cured individuals. We propose to study the effect of a covariate on the probability of cure by means of the distance correlation between this covariate and the cure indicator. The main challenge is to handle the missingness of the cure indicator of the censored individuals

Keywords: Cure model; Distance correlation; Martingale.

Acknowledgements International, Interdisciplinary and Intersectoral information and Communications technology PhD programme (3-i ICT) granted to CITIC and supported by the European Union through the Horizon 2020 research and innovation program under a Marie Skłodowska-Curie agreement (H2020-MSCA-COFUND).

References

[1] Y. Peng, B. Yu, (2021). *Cure models: methods, applications, and implementation.*. Chapman and Hall/CRC

[2] G. Székely, M. Rizzo, and N.K. Bakirov. *Measuring and testing dependence by correlation of distances.* Ann. Statist. **35** (2007) 2769-2794.

[3] B.E. Monroy-Castillo, A. Jácome, and R. Cao. *Improved distance correlation estimation.* (Submitted).

[4] D. Edelmann, T. Welchowski, and A. Benner. *A consistent version of distance covariance for right-censored survival data and its application in hypothesis testing.* Biometrics. **78** (2022) 867-879.

12th Sept
12:30-12:55
NP session I

On uniqueness of the set of k-means

Javier Cárcamo[1], Antonio Cuevas[2] and *Luis-Alberto Rodríguez*[3]

[1]Department of Mathematics, University of Basque Country, Spain;[2]Department of Mathematics, Autónoma de Madrid University, and Instituto de Ciencias Matemáticas, Consejo Superior de Investigaciones Científicas, Spain;[3]Institut für Mathematische Stochastik, Georg August University of Göttingen, Germany

We provide necessary and sufficient conditions for the uniqueness of the k-means set of a probability distribution (see [1]). This uniqueness problem is related to the choice of k: depending on the underlying distribution, some values of this parameter could lead to multiple sets of k-means, which hampers the interpretation of the results and/or the stability of the algorithms. We give a general assessment on consistency of the empirical k-means adapted to the setting of non-uniqueness and determine the asymptotic distribution of the within cluster sum of squares (see also the seminal works [2] and [3]). We also provide a statistical characterization of k-means uniqueness in terms of the asymptotic Gaussianity of the empirical WCSS. As a consequence, we derive a bootstrap test for uniqueness of the set of k-means. The results are illustrated with examples of different types of non-uniqueness that might arise. Finally, we check by simulations the performance of the proposed methodology.

Keywords: Clustering; k-means, uniqueness.

Acknowledgements We are very grateful to Professor Juan Alberto Cuesta-Albertos (Universidad de Cantabria) who read a preliminary version of this paper and made several enlightening comments. In particular, he pointed out to us the repulsion effect of k-means for distributions with support in the whole space.

References

[1] Morissette, L. and Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica *Tutorials in Quantitative Methods for Psychology*, 9(1), 15–24. 10.20982/tqmp.09.1.p015

[2] Pollard, D. (1981). Strong consistency of k-means clustering *The Annals of Statistics*, 9, 135–140. 10.1214/aos/1176345339

[3] Pollard, D. (1982). A Central Limit Theorem of k-means clustering. *The Annals of Statistics*, 10, 919–926. 10.1214/aop/1176993713

Evaluation of nonparametric machine learning algorithms. A case of study in French Covid-19 data

13th Sept
13:15-13:40
NP session III

María Luz Gámiz Pérez[1], María Dolores Martínez Miranda[1], *Germán Ernesto Silva Gómez*[2]

[1]Department of Statistics and Operations Research, University of Granada, Spain;
[2]Department of Mathematical Analysis, Statistics and Operations Research and Applied Mathematics, University of Malaga, Spain

In this study, nonparametric machine learning (ML) methods and a recent new original algorithm proposed by Gámiz et al. (2022, 2024a,b) are applied in order to monotoring and forecasting a developing pandemic. We use publicly available data from the recent Covid-19 pandemic and we compare several ML methods and the new algorithm when describing and forecasting the pandemic. One problem of this type of data is that they are aggregated and to be able to work with them some new missing data issues have to be solved. We describe and compare how ML methods and the new algorithm deal with this problem. We show that the new algorithm is able not only to explain the underlying relationships between the variables of the dataset but also to uncover crucial information that is not apparent due to the aggregated structure of the data, whereas the examined ML methods are not capable of extracting such information.

Keywords: Nonparametric Inference; Machine Learning; Missing data; Counting Processes.

References

[1] G. E. Atteia, H. A. Mengash, N. A. Samee.*Evaluation of using parametric and non-parametric machine learning algorithms for covid-19 forecasting.* International Journal of Advanced Computer Science and Applications. **12.10** 2021.

[2] M. L. Gámiz, E. Mammen, M. D. Martínez-Miranda, J. P. Nielsen. *Missing link survival analysis with applications to available pandemic data.* Computational Statistics & Data Analysis. **169** (2022) 107405.

[3] M. L. Gámiz, E. Mammen, M. D. Martínez-Miranda, J. P. Nielsen. *Low quality exposure and point processes with a view to the first phase of a pandemic.* (2024).

[4] M. L. Gámiz, E. Mammen, M. D. Martínez-Miranda, J. P. Nielsen. *Monitoring a developing pandemic with available data.* (2024).

11th Sept
10:15-10:40
FDA session III

Wavelet Spatial ICA for Wide Functional Data

Marc Vidal[1,2,3] and Ana M. Aguilera[2]

[1]Ghent University; [2]Granada University; [3]Max Planck Institute

Suppose we have observations $X_{k,i} = \Phi(t_i, s_k) + R(t_i, s_k) + \epsilon_{k,i}$ $(i = 1, \ldots, n; k = 1, \ldots, p)$, where Φ and R are random functions in a spatiotemporal domain, representing anomalies and the ground truth process, respectively, with added noise $\epsilon_{k,i}$. Here, we are interested to estimate R considering estimates of Φ when $p \ll n$, i.e., the number of temporal covariates is much larger than the number of spatial locations. A factorization into spatially independent subspaces followed by a wavelet decomposition in the temporal domain is performed to find solutions to the current problem. To estimate Φ, problems linked to the existence of optimal thresholds are addressed from a Besov regularity perspective. Two criteria are then introduced: one based on multiplicative scaling and the other on the entropic normalized distance. The efficacy of our methods is illustrated using both simulated and real EEG data, with anomalies varying in degree of sparsity in either time or space.

Keywords: Daubechies wavelet filters; FastICA; Multiresolution analysis.

Acknowledgements This research was partially supported by the Methusalem funding from the Flemish Government and the project FQM-307 of the Government of Andalusia (Spain). We also acknowledge the financial support of Agencia Estatal de Investigación, Ministerio de Ciencia e Innovación (grant number: PID2020-113961GB-I00) and the IMAG María de Maeztu grant CEX2020-001105-M/AEI/10.13039/501100011033.

References

[1] V. Bruni, I. D. Cioppa D. Vitulano. *An automatic and parameter-free information-based method for sparse representation in wavelet bases.* Math. Comput. Simul. **176** (2020) 73–95.

[2] M. Vidal, M. Rosso, A.M. Aguilera. *Bi-smoothed functional independent component analysis for EEG artifact removal.* Mathematics. **9** (2022) 1-17.

[3] A. T. Walden. *Wavelet analysis of discrete time series.* In C. Casacuberta, R. M. Miró-Roig, J. Verdera and S. Xambó-Descamps (eds), European Congress of Mathematics. Progress in Mathematics. **202** (2001) 627–641. Birkhäuser, Basel.

Kernel-based specification test for the regression function

12th Sept
12:55-13:20
NP session I

María Vidal-García[1], Ingrid Van Keilegom[2], Rosa M. Crujeiras[1] and Wenceslao González-Manteiga[1]

[1]CITMAga, Universidade de Santiago de Compostela, Spain; [2]Katholieke Universiteit Leuven, Belgium

The energy distance is a magnitude used to measure how similar two distribution functions are [3]. In some contexts, this notion is closely related to that of maximum mean discrepancy, a concept arising from the Machine Learning approach [1]. These magnitudes can be naturally applied to test hypothesis on distribution functions.

In this talk we will focus on the problem of the goodness-of-fit test for the regression function in the context of scale-location models. This problem can be equivalently posed in terms of testing the equidistribution of the constrained and unconstrained residuals [4], therefore allowing for the application of energy distance/maximum mean discrepancy in the construction of the test. The theoretical behaviour of the test can be studied using the good properties of the Reproducing Kernel Hilbert Space (RKHS) approach [2].

The performance of the proposed test will be illustrated and compared to other alternatives available in the literature via simulations.

Keywords: energy distance, maximum mean discrepancy; RKHS; scale-location model; specification test.

Acknowledgements This work is part of the R&D project PID2020-116587GB-I00 granted by MICIU/AEI/10.13039/501100011033.

References

[1] D. Sejdinovic, B. Sriperumbudur, A. Gretton, K. Fukumizu. *Equivalence of distance-based and RKHS-based statistics in hypothesis testing.* Ann. Stat. **41(5)** (2013) 2263–2291.

[2] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, G. R. Lanckriet. *Hilbert space embeddings and metrics on probability measures.* JMLR **11** (2010) 1517-1561.

[3] G. J. Székely, M. L. Rizzo. *The energy of data.* Annu. Rev. Stat. Appl. **4(1)** (2017) 447-479.

[4] I. Van Keilegom, W. González Manteiga, C. Sánchez Sellero. *Goodness-of-fit tests in parametric regression based on the estimation of the error distribution.* TEST **17** (2008) 401–415.

List of all participants

1. Acar Denizli, Nihan (Universitat Politècnica de Catalunya)
2. Ameijeiras Alonso, Jose (Universidade de Santiago de Compostela)
3. del Barrio, Eustasio (IMUVa, Universidad de Valladolid)
4. Bolón, Diego (Université Libre de Bruxelles)
5. Bongiorno, Enea G. (Università del Piemonte Orientale - Amedeo Avogadro)
6. Chacón, José E. (Universidad de Extremadura)
7. Cuevas, Antonio (Universidad Autónoma de Madrid)
8. Delicado, Pedro (Universitat Politècnica de Catalunya)
9. Diz Castro, Daniel (Universidade de Santiago de Compostela)
10. Einbeck, Jochen (Durham University)
11. Elías Fernández, Antonio (Universidad de Málaga)
12. Fanjul Hevia, Aris (Universidad de Oviedo)
13. Fernández de Marcos, Alberto (Universidad Carlos III de Madrid)
14. Freijeiro González, Laura (Universidad de Oviedo)
15. García Portugués, Eduardo (Universidad Carlos III de Madrid)
16. Hernández, Nicolás (Queen Mary University of London)
17. Hernández Amaro, Pavel (Universidad Carlos III de Madrid)
18. Jácome, M. Amalia (Universidade da Coruña)
19. Jerez López, Esther (Universidad Autónoma de Madrid)
20. Li, Zehang (Universidad Carlos III de Madrid)
21. Lillo, Rosa E. (uc3m-Santander Big Data Institute, Universidad Carlos III de Madrid)
22. Lopez Pintado, Sara (Northeastern University)
23. Martínez-Miranda, María D. (University of Granada)
24. Meilan Vila, Andrea (Universidad Carlos III de Madrid)

25. Monroy Castillo, Blanca (Universidade da Coruña)
26. Nagler, Thomas (LMU Munich, Munich Center for Machine Learning)
27. Pachón García, Cristian (Universitat Politècnica de Catalunya)
28. Ortiz, Helena (Universidad de Granada)
29. Rodríguez, Luis-Alberto (Georg August University of Göttingen)
30. Rodríguez Poo, Juan M. (Universidad de Cantabria)
31. Silva Gómez, Germán E. (Universidad de Málaga/Universidad de Granada)
32. Torrecilla, José Luis (Universidad Autónoma de Madrid)
33. Vidal, Marc (Ghent University/Universidad de Granada)
34. Vidal García, María (CITMAga, Universidade de Santiago de Compostela)