II International Workshop on Advances in Functional Data Analysis

Universidad Carlos III de Madrid

Getafe (Madrid), Spain, 23-24 March 2017

This workshop is the seventh meeting of the Functional Data Analysis (FDA) working group of the Spanish Society of Statistics and Operations Research (S.E.I.O.), http://FDA.seio.es. This meeting is co-organised with the Instituto Flores de Lemus (UC3M). There is a plenary talk by Professor Jeff Goldsmith (Columbia University) followed by shorter related presentations and a panel discussion. Also, there are two applied tutorials. The first is an introduction to FDA with the online application statfda and it will be provided by Manuel Escabias (University of Granada) and M. Carmen Aguilera (UC3M). The second one is based on variable selection in functional regression and it will be provided by Jeff Goldsmith.

Organizing Committee

- M. Carmen Aguilera (Universidad Carlos III de Madrid)
- Andrés M. Alonso Fernández (Director del Instituto Flores de Lemus, Universidad Carlos III de Madrid)

Scientific Committee (SPC)

- M. Carmen Aguilera (Universidad Carlos III de Madrid)
- Rosa Elvira Lillo (Universidad Carlos III de Madrid)
- Cristian Preda (Université Lille, France)
- Philip T. Reiss (New York University - University of Haifa)
- Juan Romo Urroz (Universidad Carlos III de Madrid)
- Mariano. J. Valderrama (Universidad de Granada)

Local Committee

- Diego Ayma (Universidad Carlos III de Madrid)
- Alba Carballo (Universidad Carlos III de Madrid)
- Antonio Elías (Universidad Carlos III de Madrid)
- Juan Carlos Laria (Universidad Carlos III de Madrid)
- Aldo L. Moreno (Universidad Carlos III de Madrid)
- Paula Navarro (Universidad de Cantabria)

Sponsors

We want to thank the sponsors of this activity:

- Sociedad de Estadística e Investigación Operativa
- Instituto Flores de Lemus, UC3M
- Departamento de Estadística, UC3M
- Universidad Carlos III de Madrid
- Facultad de Ciencias Sociales y Jurídicas, UC3M

Madrid, March 2017

PROGRAMME

PROGRAMME

Thursday, March 23:

- 09:00 - 10:00 Registration (Room: 17.2.75)
- 10:00 - 11:00 Plenary talk I (Room: 17.2.75): *Kinematic Data in Motor Control Experiments.* Jeff Goldsmith, Mailman School of Public Health, Columbia University
- 11:00 - 11:30 Coffee Break
- 11:30 - 13:30 Contributed session (Room: 17.2.75): Functional Data Analysis and Related Topics
  - Goodness-of-fit tests for the functional linear model based on randomly projected empirical processes (García Portugués, Eduardo)
  - A Modification of a Method for the Detection of Outliers in High-Dimensional Data (Navarro Esteban, Paula)
  - Ecological studies in a functional framework (Fortuna, Francesca)
  - Statistical methods for in-vivo, non-invase detection of hyperglycemic states using mm-waves spectroscopy (Moreno Oyervides, Aldo)
  - Distributional regression models including functional data. Application on continuous glucose monitoring (Espasandín, Jenifer)
  - A diagonal ARH(1)-based predictor of zero-coupon yield curve (Álvarez Liébana, Javier)
- 13:30 - 15:30 Lunch (Building 1 - Cafeteria)
- 15:30 - 17:00 Keynote I: Functional data analysis with *statfda* I (Room: 18.0A.14)
  M. Carmen Aguilera Morillo (UC3M) and Manuel Escabias Machuca (Universidad de Granada)
- 17:00 - 17:30 Coffee Break (Room: 18.0A.06)
- 17:30 - 19:00 Keynote I: Functional data analysis with *statfda* II (Room: 18.0A.14)
- 21:00 Dinner.

PROGRAMME

Friday, March 24:

- 08:30 - 09:00 Registration (Room: 18.0A.14)
- 09:00 - 10:30 Keynote II: Variable selection in functional regression I (Room: 18.0A.14)
  Jeff Goldsmith, Mailman School of Public Health, Columbia University
- 10:30 - 11:00 Coffee Break (Room: 18.0A.06)
- 11:00 - 12:30 Keynote II: Variable selection in functional regression II (Room: 18.0A.14)
  Jeff Goldsmith, Mailman School of Public Health, Columbia University
- 13:00 - 14:00 Plenary talk II (Room: Costas Goutis 10.0.23):
  Some statistical models in ecology: home range, core areas and more.
  Ricardo Fraiman, Universidad de la República, Uruguay
- 14:00 Lunch (Building 1- Cafeteria)

Plenary Talk I

Speaker Profile

Jeff Goldsmith is an assistant professor in Biostatistics at the Columbia University Mailman School of Public Health. Dr. Goldsmith joined Columbia after receiving his PhD in Biostatistics from Johns Hopkins in 2012, where his dissertation focused on statistical methods for high-dimensional structured data. Dr. Goldsmith has research interests in scientific domains including neuroimaging, physical activity monitoring using accelerometers, motion kinematics and motor learning, and urban environments. In these domains, he develops statistical methods that examine relationships between complex data structures and patient-level information; examples include using MRI scans to predict patient outcomes and examining effects of aging on daily activity patterns.

Dr. Goldsmith has published widely in the functional data analysis literature on topics including scalar-on-function regression, function-on-scalar regression, and functional principal components analysis. His work has used both Bayesian and frequentist approaches, and is implemented in public software packages. He has given over fifty presentations at conferences around the world; is the Principal Investigator for two active NIH grants; and serves as an Associate Editor for the Journal of the American Statistical Association.

Abstract

Kinematic Data in Motor Control Experiments

Invited Speaker: **Goldsmith, J.**

Experiments involving kinematic data – dense recordings of hand or finger position over time during the execution of a motion – provide deep insights into the processes underlying motor control and learning. This talk will present methods for modeling the mean and variability of kinematic data. First, we pose a function-on-scalar regression with subject-level random functional effects; parameters are jointly estimated in a Bayesian framework, and methods are applied to a study of the effects of stroke on motor control. Second, we extend the functional principal components analysis framework to allow subject and covariate effects on score variances. In a setting where the components are invariant across subjects and covariate values, this approach provides a flexible and interpretable way to explore factors that affect the variability of functional data. These methods are used to model the reduction of motion variance achieved through repetition in healthy subjects.

Plenary Talk II

Speaker
Ricardo Fraiman, Universidad de la República (Uruguay)

Abstract

Some statistical models in ecology: home range, core areas and more

Invited Speaker: **Fraiman, R.**

We study the problem of estimating the home range S of an animal from animal tracking data, an important problem in ecology. We first review briefly some of the more traditional methods reported in the bibliography. Our approach is based in modeling the movements of the animal as a trajectory of a diffusion process with reflections on the boundary of the set S. We introduce it via some real data examples. We start with the simplest model considering a reflected Brownian motion. Our approach allows us to find more flexible regions close to reality, as shown in some real data examples we provide. Next we consider a more general model in order to estimate the core area. Finally, we establish consistency and rates of convergence for some estimators of the home range S, its boundary and core areas.

Contributed session

Functional Data Analysis and Related Topics

Abstracts

Goodness-of-fit tests for the functional linear model based on randomly projected empirical processes

Speaker: **García-Portuguúes, Eduardo.**

Co-authors: Cuesta-Albertos, Juan A., Febrero-Bande, Manuel, and González-Manteiga, Wenceslao.

We consider marked empirical processes, indexed by a randomly projected functional covariate, to construct goodness-of-fit tests for the functional linear model with scalar response. The test statistics are built from continuous functionals over the projected process, resulting in computationally efficient tests that exhibit root-n convergence rate and circumvent the curse of dimensionality. The weak convergence of the process is obtained conditionally on a random direction, whilst it is proved the almost surely equivalence between the testing for significance expressed on the original and on the projected functional covariate. The computation of the test in practise involves the calibration by wild bootstrap resampling and the combination of several p-values arising from different projections by means of the false discovery rate method. The finite sample properties of the test are illustrated in a simulation study for a variety of linear models, underlying processes and alternatives. The software provided implements the tests and allows the replication of simulations and data applications. Paper available at `https://arxiv.org/abs/1701.08363`

A Modification of a Method for the Detection of Outliers in High-Dimensional Data

Speaker: **Navarro Esteban, Paula.**

Co-authors: Cuesta Albertos, Juan A., and Nieto Reyes, Alicia.

Outlier detection is an important aspect in the analysis of datasets. In the literature there exist multiple methods to detect outliers in multivariate data.

Here we present an improvement of a method proposed by the authors. The original method is based on random projections and consists in projecting the data in one-dimensional subspaces where an appropriate univariate outlier method is applied. It is important to emphasize that this method does not require the estimation of the covariance matrix. Main problem with older method was the control of its size. This problem seems to be fixed with the proposed modification. We will include some simulations showing that the proposed modification controls the size of the test and leads to improved power. It is noteworthy that both the original and the modified procedure can be applied to functional datasets.

Ecological studies in a functional framework

Speaker: **Fortuna, Francesca.**

Co-authors: Di Battista, Tonio.

Biodiversity measurement is a central topic in ecological theory as it is an important indicator for environmental assessment. To this purpose, diversity profiles provide a reliable biodiversity measure. In this context, we aim to provide a theoretical framework for evaluating diversity profiles in a functional setting. Indeed, diversity profiles may be naturally considered as functional data because they are expressed as functions of the species abundance vector in a fixed domain. As for all the functional data, in real applications, the diversity profiles are observed as a sequence of discrete data at some point of the domain. In order to reconstruct their functional form the functional data analysis approach uses different techniques such as basis function expansions or interpolation. For example, diversity profile may be suitably approximated through B-splines by placing some restrictions on the basis for preserving its convexity and non-negativeness. Although this approach well approximates the shape of the profile, it presents some limitations. Indeed, there is no best choice of the basis function and of the constraints to obtain convexity and non-negativity. Moreover, the functional approach generally works in a continuous domain while some diversity profiles lie in a discrete domain. In order to consider these aspects, a different functional approach has been proposed. Since the diversity profile is an explicit and known function of the species abundance vector, the smoothing process may be avoided, by using the known parametric form of the observations. This method allows to preserve the peculiarities of diversity profile and to analyze biodiversity

through functional instruments. In this setting, some functional statistics have been proposed for the diversity profile, such as simultaneous confidence bands for the mean function and the k-means clustering algorithm.

Statistical methods for in-vivo, non-invase detection of hyperglycemic states using mm-waves spectroscopy

Speaker: **Moreno Oyervides, Aldo.**

Co-authors: Aguilera-Morillo, M. Carmen, Durbán, María, Martín Mateos, Pedro, and Acedo, Pablo.

Hyperglycemia is a disorder in which blood glucose levels are higher than normal, leading to many medical complications. This condition is very often associated to diabetes mellitus affecting many people around the world. Nowadays, methods for diagnosis and follow up of this condition are based on enzymatic reactions, and blood samples are required. Therefore, these invasive methods have several drawbacks that affect comfort and health of patients. Because of this, the development of techniques for non-invasive blood glucose assessment has been the center of attention for many researchers for several decades. In this work, a new approach for validation and optimization of a spectroscopy system based on millimeter waves by incorporating functional data analysis is presented. To this end, a Functional Principal Component Analysis (FPCA) was performed on the measured spectral data previously smoothed by regression splines. The main results include the non-invasive detection of hyperglycemic states on living models: normoglycemic.

Distributional regression models including functional data. Application on continuous glucose monitoring

Speaker: **Espasandín-Domínguez, Jenifer.**

Co-authors: Cadarso-Suárez, Carmen, Kneib, Thomas, and Gude, Francisco.

Distributional regression models is a generic framework for performing regression analyses where every parameter of a potentially complex response distribution - and not just the mean - is related to a structured additive predictor.

In the Bayesian inference framework, Klein et al. (2015) have developed a structured additive distributional regression model. This type of model extends the use of generalized additive models (GAM, Hastie and Tibshirani,

1990) to situations in which the response distributions are non-standard, and in which not only the mean but multiple parameters are related to additive predictors via suitable link functions. Further, they allow additional flexibility by specifying structured additive predictors for each parameter of interest, and thus adjust for flexible non-linear effects of continuous covariates for which the smoothness is determined based on the data. They also allow the contemplation of spatial effects to capture unobserved spatial heterogeneity and spatial correlations, interaction terms such as varying coefficients or interaction surfaces, and cluster-specifc random effects.

The technological progress has led to the development of new measurement procedures in the form of functional data. In this work, we propose to incorporate this functional information within the framework of the distributional regression. The methodologies developed will be applied to real biomedical data, in a study of glycated haemoglobin -a test useful in the control of individuals with diabetes-. The predictor will include the results of continuous monitoring, which collects glucose measurements every 5 minutes over a week. The inclusion of glucose profiles as a predictor in distributional regression models will mark a novel advance in the study of diabetes.

A diagonal ARH(1)-based predictor of zero-coupon yield curve

Speaker: **Álvarez Liébana, Javier.**

Co-authors: Ruiz-Medina, M.D.

Zero-coupon yield curve, which shows the bonds interest rates, at a fixed time moment, in terms of specific maturity dates, is one of the key concepts in the financial and pricing theory, since its shape is the main measure of future expectations as well as the assessment of monetary policy conditions. In the context of autoregressive Hilbertian processes of order one (ARH(1) processes), where H is a real separable Hilbert space, the forecasting of the U.S. Treasury zero-coupon yield curve is displayed. In particular, a functional extension of the dynamic version of Nelson-Siegel model provided in Diebold and Li [2], will be undertaken, in order to fit the entire term structure, depending on three time-varying parameters. The derivation of a strongly-consistent, in the space of bounded linear operators, functional componentwise predictor of zero-coupon yield curve is provided, in the context of ARH(1) processes (see Álvarez-Liébana, Bosq and Ruiz-Medina [1]). Diagonal spectral decomposition of autocovariance, cross-covariance and autocorrelation operators,

in term of the eigenvectors of autocovariance operator, is regarded, such that autocorrelation operator belongs to the class of Hilbert-Schmidt operators. The diagonal design adopted plays a key role in the functional estimation and prediction, since an important dimension reduction is achieved. This fact was illustrated in the wide comparative study carried out in Álvarez-Liébana, Bosq and Ruiz-Medina [1], in comparison with alternative approaches existing in the current literature.

[1] J. Álvarez-Liébana, D. Bosq and M. D. Ruiz-Medina (2017). Asymptotic properties of a componentwise ARH(1) plug-in predictor. J. Multivariate Anal. 155, 12-34.

[2] F. X. Diebold and C. Li (2006). Forecasting the term structure of government bond yields. J. Econometrics 130, 337-364.

Keynote I

Functional data analysis with *statfda*

Speakers Profile

Manuel Escabias Machuca. Born in Valdepeñas de Jaén (Jaén). He has a degree in Science and Statistical Techniques from the Universidad de Granada (1994-1998) and is Doctor in Statistics from the same university since 2002 (with honors). He is professor in Universidad de Granada since 2009. Research lines: Stochastic prediction and modeling, functional data analysis, principal component analysis, PLS regression, functional and non-functional generalized linear models.

M. Carmen Aguilera Morillo. Born in Córdoba in 1985. Has a degree in Statistics from the Universidad de Jaén in 2006 (with honors), a degree in Science and Statistical Techniques from the Universidad de Granada in 2008 (with honors), a Master in Applied Statistics from the Universidad de Granada in 2009 and a Ph.D. in Statistics from the same university in 2013. Now, she has a tenure-track position at the department of Statistics of the Universidad Carlos III de Madrid.

Her academic and research career has been recognized by various institutions, such as the Spanish Society of Statistics and Operations Research, the Spanish Ministry of Education, the Andalusian Institute of Statistics, the Andalusian Academy of Social Sciences and Unicaja.

Research lines: penalized methods for functional data analysis (penalized PLS and principal component regression), variable selection techniques and its application to Biomedical Science and Engineering.

Co-authors: Ana M. Aguilera and Mariano J. Valderrama.

Abstract

What is Functional Data Analysis (FDA)? Functional data analysis is a statistical methodology in which the input information used to obtain the output results, is a set of curves (mathematical functions). Imaging, for example, a variable whose values depend continuously on time, it would be better to use the information of such a variable at any time point than using it only at a finite set of fixed time points. All FDA methods take advantage of using the complete curve instead of a reduced set of discrete observations that could lead to loose relevant information. This way FDA takes into account relevant features of curves (continuity, derivability, smoothness, maxima and minima, etc) to provide more accurate knowledge for making optimal decisions based on the results of statistical analysis.

How does the application *statfda* work? In spite of FDA methodologies use the complete curve to get their aims, it is technically impossible to record the mathematical function itself in the observation process of any functional data set. At most we could register discretized observations of such curves at a dense set of points of their domain (in many cases we only have sparse observations at different sampling points that could even be different for the sample units). So the process in all FDA methods and also in this application begins by uploading the discrete observations of curves. Different situations are taken into account: all curves observed at the same domain points, different curves observed at different domain points... After uploading the data, the application guides you to get the mathematical form of your curves. Finally, you will be able to select the statistical method that best suits your aim, obtaining the results of the selected analysis. The use of this application is free through the web site `http://www.statfda.com/`.

Acknowledgements: *statfda* is part of the research projects developed by the research group "Modelización y Predicción con Datos Funcionales" led by the professor Mariano Valderrama Bonnet (Universidad de Granada). Exactly, it was developed in the project P11-FQM-8068 "Métodos Estadísticos de Análisis de Datos Funcionales (ADF). Desarrollo de un Interfaz Web para su Aplicación" led by professor Ana M. Aguilera and funding by Consejería de Innovación Ciencia y Empresa, Junta de Andalucía.

Keynote II

Variable Selection in Functional Regression

Speaker Profile

Jeff Goldsmith. He is an assistant professor in Biostatistics at the Columbia University Mailman School of Public Health. Dr. Goldsmith joined Columbia after receiving his PhD in Biostatistics from Johns Hopkins in 2012, where his dissertation focused on statistical methods for high-dimensional structured data. Dr. Goldsmith has research interests in scientific domains including neuroimaging, physical activity monitoring using accelerometers, motion kinematics and motor learning, and urban environments. In these domains, he develops statistical methods that examine relationships between complex data structures and patient-level information; examples include using MRI scans to predict patient outcomes and examining effects of aging on daily activity patterns.

Dr. Goldsmith has published widely in the functional data analysis literature on topics including scalar-on-function regression, function-on-scalar regression, and functional principal components analysis. His work has used both Bayesian and frequentist approaches, and is implemented in public software packages. He has given over fifty presentations at conferences around the world; is the Principal Investigator for two active NIH grants; and serves as an Associate Editor for the Journal of the American Statistical Association.

Abstract

Variable selection techniques are now common across areas of statistics. This course will focus on variable selection in functional data analyses, including both methodological and practical aspects, and is organized in several parts.

First we will review scalar-on-function and function-on-scalar regression, focusing on spline-based representations of these models. Next, we will compare competing methodological frameworks for variable selection in these models, emphasizing penalized approaches but touching on Bayesian approaches to variable selection. Lastly, we will learn how to fit these models in R "by hand" and using packages dedicated to variable selection in functional regression.

Real data examples will be used throughout to emphasize the usefulness of these approaches.

List of participants

AGUILERA	ANA MARÍA	aaguiler@ugr.es
AGUILERA-MORILLO	M. CARMEN	mariacarmen.aguilera@uc3m.es
ALONSO FERNÁNDEZ	ANDRÉS M.	amalonso@est-econ.uc3m.es
ÁLVAREZ	MARÍA	mariamssc.ma@gmail.com
ÁLVAREZ LIÉBANA	JAVIER	javialvaliebana@ugr.es
AYMA ANZA	DIEGO	dayma@est-econ.uc3m.es
CARBALLO GONZÁLEZ	ALBA	albcarba@est-econ.uc3m.es
CARUSO	GIULIA	giulia-caruso@hotmail.it
CUESTA ALBERTOS	JUAN ANTONIO	cuestaj@unican.es
CUEVAS	ANTONIO	antonio.cuevas@uam.es
DE VICENTE MALDONADO	JAVIER	javvicen@est-econ.uc3m.es
DELGADO GOMEZ	DAVID	ddelgado@est-econ.uc3m.es
DI BATTISTA	TONIO	dibattis@unich.it
DURBÁN	MARÍA	marialuz.durban@uc3m.es
ELÍAS	ANTONIO	aelias@est-econ.uc3m.es
ESCABIAS	MANUEL	escabias@ugr.es
ESPASANDIN DOMÍNGUEZ	JENIFER	jenifer.espasandin@usc.es
FEBRERO-BANDE	MANUEL	manuel.febrero@usc.es
FORTUNA	FRANCESCA	francesca.fortuna@unich.it
GARCÍA-GURTUBAY	IGNACIO	igurtubay@gmail.com
GARCÍA-PORTUGUÉS	EDUARDO	edgarcia@est-econ.uc3m.es
GARCÍA SOUTO	JOSE A.	jsouto@ing.uc3m.es
GONZÁLEZ MANTEIGA	WENCESLAO	wenceslao.gonzalez@usc.es
GUIJARRO ROZAS	REBECA	peticiones.training@accenture.com
HALLIN	MARC	mhallin@ulb.ac.be
HERNÁNDEZ	NICOLÁS	nihernan@est-econ.uc3m.es
HERRERA	JORGE	joherrer@est-econ.uc3m.es
JUSTEL EUSEBIO	ANA	ana.justel@uam.es
LARIA DE LA CRUZ	JUAN CARLOS	jlaria@est-econ.uc3m.es
LILLO	ROSA. E.	lillo@est-econ.uc3m.es
MÉNDEZ CIVIETA	ÁLVARO	peticiones.training@accenture.com
MORENO OYERVIDES	ALDO	aldmoren@ing.uc3m.es
NAVARRO ESTEBAN	PAULA	navarrop@unican.es
NGUYEN	HOANG	hoang.uc3m@gmail.com
NIETO REYES	ALICIA	alica.nieto@unican.es
RODRÍGUEZ PAINO	CARMEN	peticiones.training@accenture.com
RUIZ-CASTRO	JUAN ELOY	jeloy@ugr.es
SGUERA	CARLO	csguera@est-econ.uc3m.es
TORRECILLA	JOSÉ L.	joseluis.torrecilla@uc3m.es
VALDERRAMA BONNET	MARIANO JOSÉ	valderra@ugr.es

Planos de los edificios

1. Cafetería - autoservicio
2. Edificio de servicios
3. Edificio Decanato
4. Edificio Gómez de la Serna
5. Edificio Giner de los Ríos
6. Edificio Normante
7. Edificio Foronda
8. Edificio Rectorado
9. Edificio Adolfo Posada
10. Edificio Campomanes
11. Edificio Luis Vives
12. Edificio María Moliner
13. Centro Deportivo Ignacio Pinedo
14. Edificio Concepción Arenal
15. Edificio López Aranguren
16. Cafetería
17. Edificio Ortega y Gasset
18. Edificio Carmen Martín Gaite
19. C. Deportivo Seve Ballesteros
20. Residencia de estudiantes Gregorio Peces Barba
21. Residencia de estudiantes Fernando de los Ríos