I International Workshop on Advances in Functional Data Analysis

University Carlos III de Madrid

Getafe (Madrid), Spain, 11-12 November 2015

This workshop is the fifth meeting of the FDA (Functional Data Analysis) working group of the Spanish Society of Statistics and Operational Research (S.E.I.O.) (http://FDA.seio.es). There is a plenary talk by Professor Philip Reiss followed by shorter related presentations and a panel discussion. The scientific meeting gives researchers opportunities to informally discuss novel, controversial or educational topics with respect to new methodologies and applications in Functional Data Analysis. There will be a keynote delivered by Carlos Gil Bellosta introducing tools to get more out of R with massive data.

Organizing Committee

- Rosa Elvira Lillo (Universidad Carlos III de Madrid)
- M. Carmen Aguilera (Universidad Carlos III de Madrid)

Scientific Committee (SPC)

- Rosa Elvira Lillo (Universidad Carlos III de Madrid)
- M. Carmen Aguilera (Universidad Carlos III de Madrid)
- Ana. M. Aguilera (Universidad de Granada)
- Pedro Delicado (Universitat Politècnica de Catalunya)
- Hans. G. Müller (University of California)
- Cristian Preda (Université Lille 1)
- Juan Romo (Universidad Carlos III de Madrid)
- Mariano. J. Valderrama (Universidad de Granada)

Local Committee

- Elisa Cabana (Universidad Carlos III de Madrid)
- Alba Carballo González (Universidad Carlos III de Madrid)
- Antonio Elías Fernández (Universidad Carlos III de Madrid)
- María Jesús Gisbert Francés (Universidad Carlos III de Madrid)
- Janeth Carolina Rendón Aguirre (Universidad Carlos III de Madrid)

Sponsors
We want to thank the sponsors of this activity:

- Sociedad de Estadística e Investigación Operativa (S.E.I.O.)
- Universidad Carlos III de Madrid
- Facultad de Ciencias Sociales y Jurídicas (UC3M)
- Departamento de Estadística (UC3M)

Madrid, November de 2015

PROGRAMME

PROGRAMME

Wednesday, November 11:

- 09:00 - 10:00 Registration (Room: 17.2.75)
- 10:00 - 11:00 Plenary talk (Room: 17.2.75):
  Flexible penalized regression for functional data... and other complex data objects. Philip T. Reiss, New York University & University of Haifa in Israel
- 11:00 - 11:30 Coffee Break
- 11:30 - 13:30 Contributed session (Room: 17.2.75): Functional Data Analysis and Related Topics
  - An online application for applying functional data analysis without expert knowledge (Escabias, Manuel)
  - Outlier Detection in High-Dimensional Data Using Random Projections (Navarro Esteban, Paula)
  - Variable selection based on reproducing kernels (Torrecilla, José Luis)
  - Robust clustering for functional data (Rivera-García, Diego)
  - Applications of the functional Mahalanobis semi-distance (Gisbert Francés, M. Jesús)
  - Forecasting with generalized additive models (Carballo González, Alba)
  - Robust regression based on depth measures for the fMRI problem (Cabana, Elisa)
  - The development of the Scoring models process (Ramos de Alvaro, José Ignacio)
- 13:30 - 16:00 Lunch (Building 1 - Cafeteria)
- 16:00 - 18:00 Discussion session: Professor Reiss and young members of the FDA working group (Room: Costas Goutis - 10.0.23)
- 20:30 Dinner.

PROGRAMME

Thursday, November 12:

- 09:00 - 10:00 Registration (Room: 18.0A.02)
- 10:00 - 12:00 Keynote: Working in R with massive data I (Room: 18.0A.02)
- 12:00 - 12:30 Coffee Break
- 12:30 - 14:00 Keynote: Working in R with massive data II (Room: 18.0A.02)
- 14:00 Lunch (Building 1- Cafeteria)

Plenary Talk

Speaker Profile

Philip Reiss received his Ph.D. in Biostatistics at Columbia University in 2006. Since then he has been with the Division of Biostatistics in the Department of Child and Adolescent Psychiatry at New York University School of Medicine in New York City. He is now an Associate Professor in that department, with secondary affiliations at the Department of Population Health and the Nathan S. Kline Institute for Psychiatric Research. Prof. Reiss's research focuses primarily on semiparametric regression and functional data analysis, with emphasis on methods for analyzing psychiatric neuroimaging data. Currently he is visiting the Department of Statistics at the University of Haifa in Israel.

Abstract

Flexible penalized regression for functional data... and other complex data objects

Invited Speaker: **T. Reiss, Philip.**

In the 1990s, Ramsay and Silverman proposed a set of roughness penalty approaches for regression with functional predictors only (later termed "scalar-on-function" regression), functional responses only ("function-on-scalar" regression), or both ("function-on-function" regression). More recently, thanks to the connection between roughness penalties and mixed models, as well as advances in automatic smoothing parameter selection, penalized approaches to functional regression have become much more flexible: they can now readily incorporate multiple predictors, generalized linear outcomes, and random effects. This talk will describe the strategy for flexible penalized regression with functional data that is implemented in the "refund" package for R. We will also show how the same paradigm can be extended from scalar-on-function regression to settings in which the predictors are more general data objects (images, graphs, etc.), equipped with either a distance measure or a similarity measure such as a kernel. The resulting method can be viewed as a simple but powerful new technique for nonparametric functional regression, or for kernel learning.

Contributed session

Functional Data Analysis and Related Topics

Abstracts

An online application for applying functional data analysis without expert knowledge

Speaker: **Escabias, Manuel.**

Co-authors: Aguilera, Ana, Aguilera-Morillo, M. Carmen and Valderrama, Mariano.

It is well known that Functional Data Analysis has sparked the interest of Statisticians in the last years. More and more researchers from different fields have been also attracted by this methodology in order to analyze curves from a functional point of view. It is difficult for these researchers to find the accurate software to make a Functional Data Analysis of their curves, because these methods are not implemented in the most usual commercial software (SPSS, STATA, SAS...) that works with software menus and mouse movements. Moreover the use of the best developed software for Functional Data Analysis requires a deep knowledge of the methodology and the programming language. This software is available in S-Plus, R and Matlab. For most of applied researchers of field different than Statistics this knowledge is not usual. With the objective of bringing Functional Data Analysis to applied researchers, we have developed a Web based application that let these researchers to apply Functional Data Analysis by using a friendly graphic interface. The application works by uploading the matrix of discrete observations of the curves and by choosing among different suggested options for the functional data analysis (type of basis, dimension, observations knots,...).

Acknowledgments: Funded by Project P11-FQM-8068 from Consejería de Innovación, Ciencia y Empresa. Junta de Andalucía, Spain.

Outlier Detection in High-Dimensional Data Using Random Projections

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, but most of them require to estimate the matrix of covariance. The higher the dimension, the more complex the estimation of the matrix due to the sparsity of high dimensions. In order to avoid estimating this matrix, a procedure to detect outliers in Gaussian multivariate data based on random projections is proposed. It consists in projecting the data in one-dimensional subspaces where an appropriate univariate outlier method is applied. The required number of projections is also provided. It is noteworthy that this procedure can be applied to functional datasets. To illustrate the method that we introduce, simulated and real datasets are studied.

Variable selection based on reproducing kernels

Speaker: **Torrecilla, José Luis.**

Co-authors: Berrendero, José R. and Cuevas, Antonio.

Variable selection techniques have become a popular tool for dimension reduction with an easy interpretation. However, we are still far from getting a standard in the functional classification framework. Here we propose a new functional-motivated variable selection methodology (RK-VS). This method appears as a direct consequence of looking at the functional classification problem from an RKHS (Reproducing Kernel Hilbert Space) point of view. In this context, under a general Gaussian model and a sparsity assumption, the optimal rules turn out to depend on a finite number of variables. These variables can be selected by maximizing the Mahalanobis distance between the finite-dimensional projections of the class means. Our RK-VS method is an iterative approximation to this. This is an easy-to-interpret and fast methodology which allows for easily adding extra information about the model. The empirical performance of RK-VS is extremely good when the considered problems fit the assumed model but it turns out to be also quite robust against partial departures from the hypotheses, typically leading to very good results in general problems.

Robust clustering for functional data

Speaker: **Rivera-García, Diego.**

Co-authors: García-Escudero, L.A. and Mayo-Iscar, A.

Many algorithms for clustering analysis when the data are curves or functions have been proposed recently. However the presence of contamination in the data can influence the performance of most of these clustering techniques. Therefore, it would be interesting to get available tools for robustifying clustering algorithms.

In this work, we propose a robust clustering method based on approximate coordinates obtained by applying functional principal components. This robustness is based on the joint application of trimming, for reducing the effect of contaminated observations, and constraints on the variances, for avoiding spurious clusters in the solution. The proposed method was evaluated through a simulation study, which showed an improved performance when compared with other recent methods for functional data clustering.

Applications of the functional Mahalanobis semi-distance

Speaker: **Gisbert Francés, M. Jesús.**

Co-authors: Lillo, Rosa and Galeano, Pedro.

The use of distances in multivariate analysis is very common in many different problems as classifications, clustering, hypothesis testing or outlier detection, among others. In particular, Mahalanobis distance has developed an important and fundamental role in multivariate statistics. However, classical techniques designed for multivariate analysis are no applicable for functional data. Rosa Lillo, Pedro Galeano and Esdras Joseph adapted the Mahalanobis distance for functional data defining the functional Mahalanobis semi-distance. They also applied this semi-distance for classification and hypothesis testing. Following this research line, my objective is to continue surveying the usefulness of the functional Mahalanobis semi-distance in other statistical problems based on distances. Specifically, Rosa Lillo, Pedro Galeano and I are now working on outlier detection. In the future, we would like to address the clustering problem.

Forecasting with generalized additive models

Speaker: **Carballo González, Alba.**

Co-authors: Durbán, María and Lee, Dae-Jin.

Additive models are a class of non-parametric regression methods which have been found widespread applications in practice, they can be used to model in many areas such as electricity load or mortality rate, where, due to the kind of data and to the information that can be known, it is important to forecast future observations. Our main objective is the development of a unified approach for forecasting with generalized additive models.

One of the main assumptions of additive models is that the effect of covariates on the dependent variable follows an additive form. We model the separate effects by smoothing penalized splines, i.e. using a regression basis and modifying the likelihood function by adding a penalty term over adjacent regression coefficients to control the smoothness of the fit. The general framework to forecast new observations is to extend the regression basis used for fitting and the penalty to control the smoothness.

We intend to work in several topics, for instance: forecasting with correlated data and in more than one dimension, applying the proposed methods in demography and insurance industry to forecast mortality tables, or in environmental sciences, and epidemiology to extrapolate spatio-temporal data. We would also like to explore other areas in which these methods have a natural application such as Functional regression forecasting.

Robust regression based on depth measures for the fMRI problem

Speaker: **Cabana Garceran del Vall, Elisa.**

Co-authors: Lillo, Rosa.

Functional Magnetic Resonance Imaging (fMRI) is one of the top techniques within the neuroimaging field. The aim of fMRI data analysis is to determine which regions of the brain are either activated or inactivated with respect to an experimental design. There exists many statistical methodologies to approach the analysis but it is obvious the need of finding a balance between interpretability, computational cost and robustness, due to the main characteristics of these massive and complicated data.

The proposal that Rosa Lillo and I (together with the collaboration of the Medical Imaging Laboratory, Unit of Experimental Surgery and Medicine,

General Universitary Hospital, Gregorio Marañón), are currently investigating is a new robust and computationally efficient approach, based on data depth, to address the fMRI problem. Until now, this approach provides good empirical results when comparing with other existing methodologies.

The development of the Scoring models process

Speaker: **Ramos de Alvaro, José Ignacio.**

Co-authors: Risk Department, axesor.

axesor is a leader in providing business, financial and marketing information. This has enabled it to become the first and only Spanish company registered as a European rating agency. To do this, axesor has developed a wide variety of services and tools that facilitate the decision-making of our clients. One of the main areas of axesor's work is to analyse commercial risk management. To perfom this task, Scoring Models have been developed, which allow the automated evaluation of a large number of companies.

To date, in order to make these Scoring Models, information is collected from different public organizations. This information refers to the economic, financial, behavioural and qualitative structure of the company. However, the business environment of recent years in combination with the advances in the processing of large volumes of information, both in terms of processing capacity and modeling, have created the need to develop more sophisticated tools.

This situation has led axesor to seek new information sources with which to enrich its data warehouse, developing a system that allows us to share highly-detailed B2B information. Thus, processing and exploring this new information has become a great challenge. In this sense, the statistical software R allows us to handle and process a large amount of data. In addition, it allows us to develop data libraries adapted to our methodological and quality requirements.

Keynote

Working in R with massive data

Speaker Profile

Carlos J. Gil Bellosta is a mathematician with over ten years of experience as freelance statistical consultant in companies such as Barclays, BBVA and Banco de Santander in the banking sector or eBay in the high tech industry. He is an R enthusiast, has authored several packages in R, and is the current president of the association of Spanish users of R (Comunidad R Hispano).

Abstract

Invited Speaker: **Gil Bellosta, Carlos. J.**

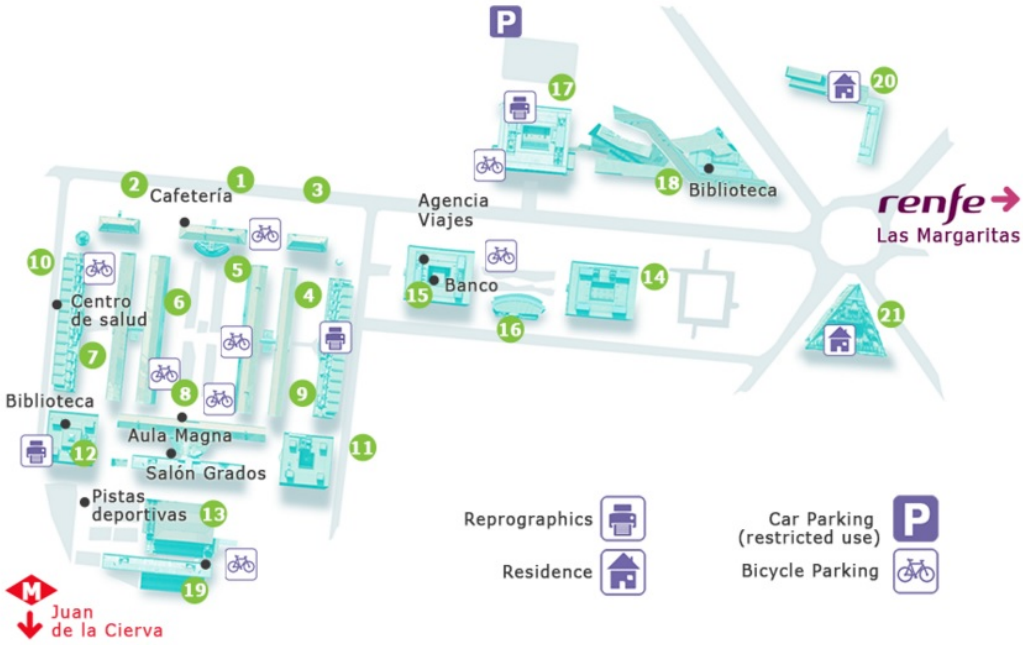
The course is an introduction to R tools for efficient big data processing and analysis. It consists of three different parts: parallelization, the data.table package and RHadoop. The first part is a short introduction to process parallelization in R.

In the second part, we will explore the data.table package, which implements a modified version of traditional data.frames in R and allows for rapid processing of large collections of data. However, these are still datasets fitting in a single machine.

As an extension of the first two parts, we will learn how to process datasets beyond the single machine limit using RHadoop. We will focus in the statistical analysis of data and the challenges posed by the need to parallelize statistical algorithms.

List of participants

ANA MARÍA	AGUILERA	aaguiler@ugr.es
M CARMEN	AGUILERA MORILLO	maguiler@est-econ.uc3m.es
LAURA	ANTÓN SÁNCHEZ	l.anton-sanchez@upm.es
ANA	ARRIBAS GIL	ana.arribas@uc3m.es
JOSÉ RAMÓN	BERRENDERO	joser.berrendero@uam.es
BEATRIZ	BUENO LARRAZ	beatriz.bueno@uam.es
ELISA	CABANA GARCERAN DEL VALL	ecabana@est-econ.uc3m.es
ALBA	CARBALLO GONZÁLEZ	albcarba@est-econ.uc3m.es
IRENE	CÓRDOBA-SÁNCHEZ	icordoba@fi.upm.es
JUAN ANTONIO	CUESTA ALBERTOS	cuestaj@unican.es
ANTONIO	CUEVAS	antonio.cuevas@uam.es
MARÍA	DURBÁN	marialuz.durban@uc3m.es
ANTONIO	ELÍAS FERNÁNDEZ	aelias@est-econ.uc3m.es
MANUEL	ESCABIAS	escabias@ugr.es
SARA	FERNÁNDEZ	saraback83@hotmail.com
PEDRO	GALEANO	pedro.galeano@uc3m.es
RAUL	GALLARDO	rgallardo@tradingsystemclub.com
CARLOS	GIL BELLOSTA	cgb@datanalytics.com
M JESÚS	GISBERT FRANCÉS	mgisbert@est-econ.uc3m.es
GINETTE	LAFIT	glafit@est-econ.uc3m.es
DAE-JIN	LEE	dlee@bcamath.org
IGNACIO	LEGUEY	ig.leguey@upm.es
FLORIAN	LEITNER	florian.leitner@gmail.com
FEDERICO	LIBERATORE	fliberatore@gmail.com
ROSA ELVIRA	LILLO	lillo@est-econ.uc3m.es
LING	LIU	lliu@est-econ.uc3m.es
PAULA	NAVARRO ESTEBAN	paula.navarro@unican.es
JAVIER	NOGALES	fjnm@est-econ.uc3m.es
ALBERTO	OGBECHIE	a.ogbechie@upm.es
LARA	QUIJANO	laraquijano@hotmail.com
JOSE IGNACIO	RAMOS DE ALVARO	jramos@axesor.es
PHILIP	REISS	Phil.Reiss@nyumc.org
CAROLINA	RENDÓN AGUIRRE	jrendon@est-econ.uc3m.es
DIEGO	RIVERA GARCÍA	driver@cimat.mx
CARLO	SGUERA	csguera@est-econ.uc3m.es
NOELIA	SIERRA PÉREZ	nsierra@axesor.es
JOSE LUIS	TORRECILLA	joseluis.torrecilla@uam.es
MARIANO J.	VALDERRAMA	valderra@ugr.es
GHERARDO	VARANDO	gherardo.varando@upm.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