PROGRAMME AND ABSTRACTS

Final CRoNoS meeting and 2nd workshop and training school on

Multivariate Data Analysis and Software (CRoNoS MDA 2019)

http://cmstatistics.org/CRONOSMDA2019/

Poseidonia Beach Hotel, Limassol, Cyprus
14-16 April 2019

Dear Friends and Colleagues,

We welcome you warmly to Limassol for the final CRoNoS meeting and the 2nd workshop and training school on *Multivariate Data Analysis and Software* (CRoNoS & MDA 2019). These events are organized in collaboration with the COST Action *Computationally-intensive and RObust analysis of NOn-Standard data* (CRoNoS), the Cyprus University of Technology and the Frederick University. CRoNoS is a network of over 80 European researchers spanning computing, statistics, machine learning, and mathematics. Their aim is to develop new models, methods and efficient, numerically stable, and well-conditioned robust strategies to improve knowledge extraction from non-perfect and non-standard datasets. More information is available at http://www.cronosaction.com/.

The CRoNoS & MDA 2019 programme consists of a course of approximately 19 hours complemented with 3 plenary talks, about 30 sessions and 145 presentations. The CRoNos Chairs have endeavoured to provide a balanced and stimulating programme that will appeal to the diverse interests of the participants. It is hoped that the conference venue will provide the appropriate environment to enhance your contacts and to establish new ones. We acknowledge the support of our hosts and sponsors, and particularly the Cyprus University of Technology, the Frederick University and the COST office.

The Elsevier journal, Econometrics and Statistics (EcoSta) is related to the CRoNoS Action. The EcoSta is the official journal of the networks of Computational and Financial Econometrics (CFEnetwork) and of Computational and Methodological Statistics (CMStatistics). It publishes research papers in all aspects of econometrics and statistics and it comprises two sections, namely, Part A: Econometrics and Part B: Statistics. The participants are encouraged to submit their papers to special or regular peer-reviewed issues of EcoSta.

We wish you a productive, stimulating workshop/course and a memorable stay in Limassol.

The CRoNoS & MDA 2019 Chairs
Ana Colubi, Elena Fernandez-Iglesias, M. Brigida Ferraro and Erricos J. Kontoghiorghes.

SOCIAL EVENTS

- *The coffee breaks* will take place at the Foyer of the Mezzanine of the Poseidonia Beach Hotel. Participants must have their conference badge in order to attend the coffee breaks.
- *Lunches* will take place at the Poseidon Restaurant of the Poseidonia Beach Hotel. Lunches are optional and registration is required. Information about the purchased lunch tickets is embedded in the QR code on the conference badge. Participants must have their conference badge in order to attend the lunch each day.
- *Welcome Reception, Sunday 14th of April 2019, from 19:30-21:00.* The Welcome Reception is open to all registrants and accompanying persons who have purchased a reception ticket. It will take place at the Poseidon terrace of the Poseidonia Beach Hotel. Participants must bring their conference badge in order to attend the reception. Preregistration is required due to health and safety reasons.
- *Workshop and Spring Course Dinner, Monday 15th of April 2019, from 20:30 to 23:00.* The conference dinner is optional and registration is required. It will take place at the Karatello Tavern (24 Vasilissis Str. Limassol 4533 - see map at page V). It is 20 mins by bus or taxi from the conference venue. Information about the purchased lunch tickets is embedded in the QR code on the conference badge. Participants must bring their conference badge in order to attend the conference dinner.
- *Closing Dinner, Tuesday 16th of April 2019, from 20:00 to 22:30.* The closing dinner is optional and registration is required. It will take place at the Kyrenia Nautical Club (Amathountos 1, Mouttagiaka 4531 - see map at page V). It is 10 mins walk from the conference venue. Information about the purchased lunch tickets is embedded in the QR code on the conference badge. Participants must bring their badge in order to attend the dinner.

GENERAL INFORMATION

Address of the venue

- 25, Amathus Avenue, Agios Tychonas, 4532 Limassol, Cyprus.

Registration

The registration will be open at the Foyer of the Mezzanine from 08:30 to 09:30 and during the coffee breaks. Please note that those selected the **eco-registration** will receive only a badge and no other material. Thus, they should copy the programme and book of abstracts in their electronic devises.

Lecture rooms

The CRoNoS and Workshop sessions will take place at the Room Neptune, located at the Ground Floor of the Poseidonia Beach Hotel, and Triton 1+2 and Business Center located at the Mezzanine. The course will take place at the Room Triton 3 located at the Mezzanine as well.

Presentation instructions

The lecture rooms will be equipped with a mini-PC and a computer projector. The session chairs should obtain copies of the talks on a USB stick before the session starts (use the lecture room as the meeting place), or obtain the talks by email prior to the start of the conference. Presenters must provide the session chair with the files for the presentation in PDF (Acrobat) format on a USB memory stick. This must be done at least ten minutes before each session. Chairs are requested to keep the sessions on schedule. Papers should be presented in the order they are listed in the programme for the convenience of attendees who may wish to go to other rooms mid-session to hear particular papers. In the case of a presenter not attending, please use the extra time for a break or a discussion so that the remaining papers stay on schedule.

Posters

The poster sessions will take place at the Room Business Center located at the Mezzanine. The posters should be displayed only during their assigned session. The authors will be responsible for placing the posters in the poster panel displays and removing them after the session. The maximum size of the poster is A0.

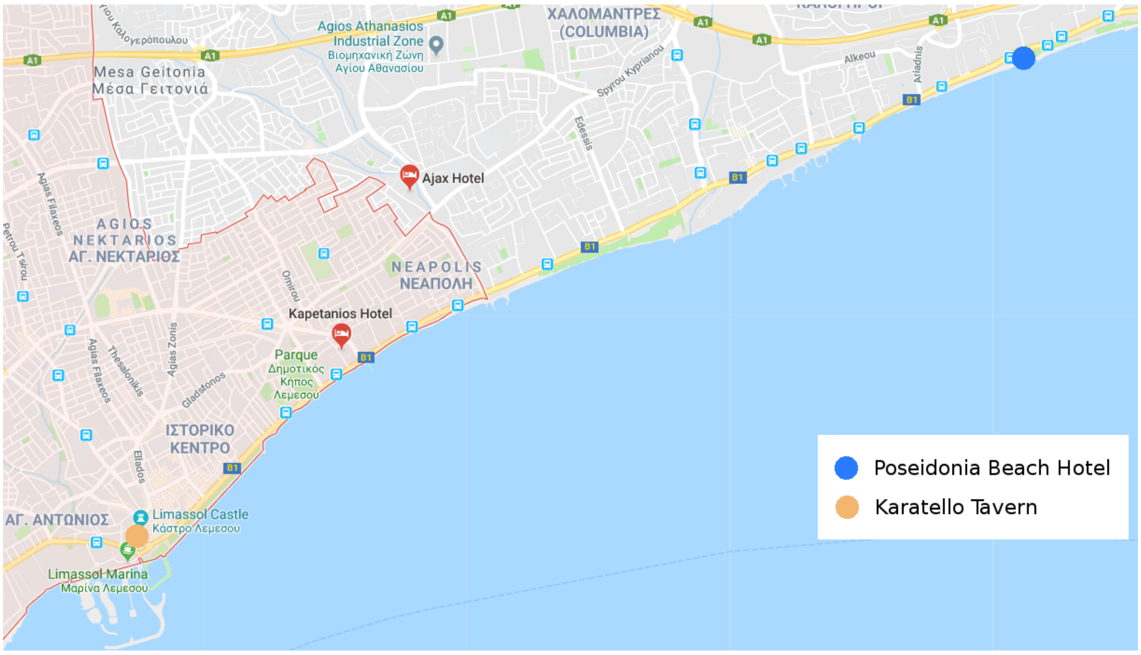
Internet Connection

There will be wireless Internet connection at the venue. You will need to have your own laptop in order to connect to the Internet. The login and password will be displayed on the announcement board by the registration desk.

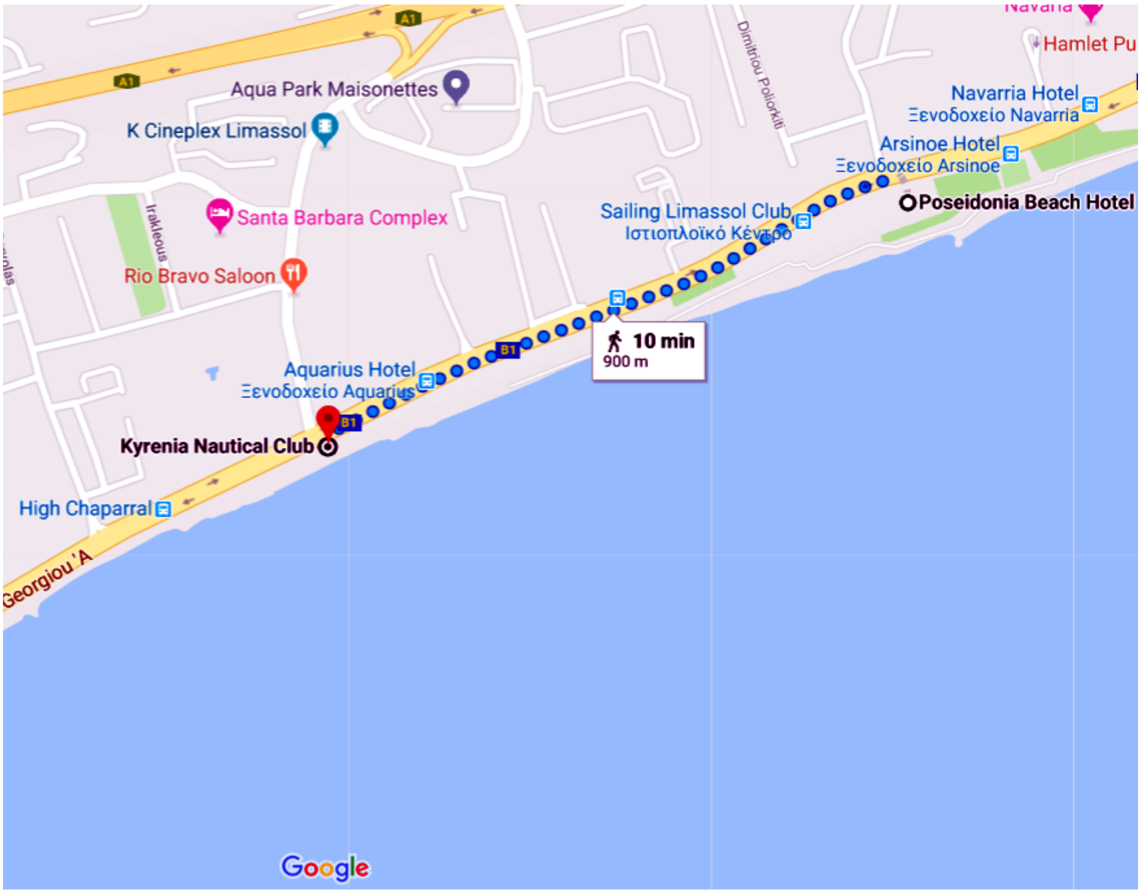
SCIENTIFIC PROGRAMME

2019-04-14	2019-04-15	2019-04-16
Opening, 09:20 - 09:30	**G** 08:50 - 10:00	**L** 08:50 - 10:20
A - Keynote 09:30 - 10:20	**Coffee Break** 10:00 - 10:30	**Coffee Break** 10:20 - 10:50
Coffee Break 10:20 - 10:50	**H - Keynote** 10:30 - 11:20	**M - Keynote** 10:50 - 11:40
B 10:50 - 12:50	**I** 11:30 - 13:00	**N** 11:50 - 13:00
Lunch Break 12:50 - 14:20	**Lunch Break** 13:00 - 14:30	**Lunch Break** 13:00 - 14:30
C 14:20 - 15:50	**J** 14:30 - 16:30	**O** 14:30 - 16:00
Coffee Break 15:50 - 16:20	**Coffee Break** 16:30 - 17:00	**Coffee Break** 16:00 - 16:30
D 16:20 - 17:30	**K** 17:00 - 18:30	**P** 16:30 - 18:30
E 17:40 - 18:50	**Conference Dinner** 20:30 - 22:30	**Closing Dinner** 20:00 - 22:00
F 18:50 - 19:20		
Welcome Reception 19:30 - 21:00		

Map of the venue and nearby area



Route for the closing dinner



PUBLICATION OUTLETS

Econometrics and Statistics (EcoSta)

`http://www.elsevier.com/locate/ecosta`

Econometrics and Statistics (EcoSta), published by Elsevier, is the official journal of the networks Computational and Financial Econometrics and Computational and Methodological Statistics. It publishes research papers in all aspects of econometrics and statistics and comprises two sections:

Part A: Econometrics. Emphasis is given to methodological and theoretical papers containing substantial econometrics derivations or showing a potential of a significant impact in the broad area of econometrics. Topics of interest include the estimation of econometric models and associated inference, model selection, panel data, measurement error, Bayesian methods, and time series analyses. Simulations are considered when they involve an original methodology. Innovative papers in financial econometrics and its applications are considered. The covered topics include portfolio allocation, option pricing, quantitative risk management, systemic risk and market microstructure. Interest is focused as well on well-founded applied econometric studies that demonstrate the practicality of new procedures and models. Such studies should involve the rigorous application of statistical techniques, including estimation, inference and forecasting. Topics include volatility and risk, credit risk, pricing models, portfolio management, and emerging markets. Innovative contributions in empirical finance and financial data analysis that use advanced statistical methods are encouraged. The results of the submissions should be replicable. Applications consisting only of routine calculations are not of interest to the journal.

Part B: Statistics. Papers providing important original contributions to methodological statistics inspired in applications are considered for this section. Papers dealing, directly or indirectly, with computational and technical elements are particularly encouraged. These cover developments concerning issues of high-dimensionality, re-sampling, dependence, robustness, filtering, and, in general, the interaction of mathematical methods, numerical implementations and the extra burden of analysing large and/or complex datasets with such methods in different areas such as medicine, epidemiology, biology, psychology, climatology and communication. Innovative algorithmic developments are also of interest, as are the computer programs and the computational environments that implement them as a complement.

The journal consists, preponderantly, of original research. Occasionally, review and short papers from experts are published, which may be accompanied by discussions. Special issues and sections within important areas of research are occasionally published. The journal publishes as a supplement the Annals of Computational and Financial Econometrics.

Call For Papers Econometrics and Statistics (EcoSta)

`http://www.elsevier.com/locate/ecosta`

Papers containing novel components in econometrics and statistics are encouraged to be submitted for publication in special peer-reviewed, or regular issues of the new Elsevier journal Econometrics and Statistics (EcoSta) and its supplement Annals of Computational and Financial Econometrics.

Papers should be submitted using the Elsevier Electronic Submission tool EES: http://ees.elsevier.com/ecosta (in the EES please select type of article “EcoSta conference”). For further information please consult http://www.cfenetwork.org or http://www.cmstatistics.org.

Contents

Sunday 14.04.2019 09:30 - 10:20 Room: Neptune Chair: Elvezio Ronchetti Keynote talk 1

Stepwise regression ensembling

Speaker: **Stefan Van Aelst, University of Leuven, Belgium**

Ensembling is a powerful approach to model complex relations in high-dimensional data and yield accurate predictions. However, it is not obvious which models are best combined in an ensemble. Standard approaches are to use a predefined set of models or to use some sort of randomness to build models. We propose a data-driven approach in which the different models are grown simultaneously. The candidate variables are added to the models in a stepwise manner. Variables are combined in a single model if they work well together. Otherwise they are assigned to different models. The resulting ensemble likely overfits the data. Therefore, we regularize each of the models using lasso or elastic net penalties. These models are then combined in an ensemble to obtain the final fit. We show the performance of the method using simulations and real data examples and compare it to existing approaches.

Monday 15.04.2019 10:30 - 11:20 Room: Neptune Chair: Jochen Einbeck Keynote talk 2

How good is your selected subgroup?

Speaker: **Xuming He, University of Michigan Ann Arbor, United States**

Subgroup analysis is frequently used to account for the treatment effect heterogeneity in clinical trials. When a treatment is seen marginally effective for the population of the original study, it is tempting to consider post hoc subgroup identification. When a highly promising subgroup is selected this way, serious questions have to be asked about the potential risks and rewards of the subgroup pursuit. We will discuss factors that have direct impacts on the credibility of subgroup pursuit, and then propose a model-free approach to quantify how likely the promise of the selected subgroup is a statistical artifact, and how good the selected subgroup really is. The proposed quantitative analysis of subgroup pursuit can help inform better decisions about any selected subgroup in clinical trials.

Tuesday 16.04.2019 10:50 - 11:40 Room: Neptune Chair: Aurore Delaigle Keynote talk 3

Object oriented data analysis

Speaker: **Steve Marron, University of North Carolina at Chapel Hill, United States**

The rapid change in computational capabilities has made Big Data a major modern statistical challenge. Less well understood is the rise of Complex Data as a perhaps greater challenge. Object Oriented Data Analysis (OODA) is a framework for addressing this, in particular providing a general approach to the definition, representation, visualization and analysis of Complex Data. The notion of OODA generally guides data analysis, through providing a useful terminology for interdisciplinary discussion of the many choices typically needed in modern complex data analyses. The main ideas are illustrated via a survey of a number of approaches which integrate differential geometry and Bayesian statistics, yielding powerful image segmentations.

Sunday 14.04.2019 10:50 - 12:50 Parallel Session B – CRONOSMDA2019

CI013 Room Triton 3 SPRING COURSE SESSION I Chair: Karel Hron

C0151: **Applied compositional data analysis**
Presenter: **Karel Hron**, Palacky University, Czech Republic

Compositional data are multivariate observations that carry relative information. They are measured in units like proportions, percentages, mg/l, mg/kg, ppm, and so on, i.e., as data that might obey (or not) a constant sum of components. Due to their specific features, the statistical analysis of compositional data must obey the geometry of the simplex sample space. In order to enable processing of compositional data using standard statistical methods, compositions can be conveniently expressed by means of real vectors of logratio coordinates. Their meaningful interpretability is of primary importance in practice. Aim of the course is to introduce the logratio methodology of compositional data together with a wide range of its possible applications. The first part of the course will be devoted to theoretical aspects of the methodology including principles of compositional data analysis, geometrical representation of compositions, construction of logratio coordinates and their interpretability. In the second part exploratory data analysis including visualization will be presented, followed by concrete popular statistical methods, e.g. correlation and regression analysis, or principal component analysis, and even methods for processing of high-dimensional data adapted within the logratio methodology. Also robust counterparts to some of these methods will be discussed. Numerical examples will be presented using the package robCompositions of the statistical software R.

CO029 Room Neptune SOME ADVANCES IN FUNCTIONAL STATISTICS: THEORY AND APPLICATIONS Chair: Enea Bongiorno

C0200: **Prior specification for the functional linear regression model**
Presenter: **Maria Franco Villoria**, University of Torino, Italy
Co-authors: Massimo Ventrucci, Haavard Rue

The functional linear regression model is considered as a varying coefficient model in a Bayesian setting. Within this framework, the functional coefficient can be seen as a smooth stochastic process such as a random walk. Elicitation of prior distributions that avoid overfitting is investigated following the recently introduced penalized complexity prior framework.

C0216: **Testing the equality of a large number of means of Hilbert data**
Presenter: **Maria Dolores Jimenez-Gamero**, Universidad de Sevilla, Spain
Co-authors: Alba Franco-Pereira

Given k independent samples of data taking values in an Hilbert space, the problem of testing for the equality of their means is considered. In contrast to the classical setting, where k is fixed and the sample size from each population increases without bound, k is assumed to be large and the size of each sample is either bounded or small in comparison to k. The asymptotic distribution of the considered test statistic is studied under the null hypothesis of equality of the k means as well as under alternatives. As special cases, functional data and finite dimensional data are considered.

C0222: **Tests and confidence regions for incompletely observed functional data**
Presenter: **David Kraus**, Masaryk University, Czech Republic

Methods are studied for the analysis of functional data under partial observation, by which we mean situations, where each functional variable may be observed only on a subset of the domain while no information about the function is available on the complement. Interestingly, some essential methods, such as K-sample tests of equal means or covariances and confidence intervals for eigenvalues and eigenfunctions, that are well established for completely observed curves, are lacking under the incomplete observation regime. The only currently available approach, in which incomplete curves are omitted, is clearly suboptimal and even infeasible, if there are no complete curves. We study methods that use all curve fragments and do not even require any complete curves. The principal difficulty in the practical implementation is the impossibility to perform dimension reduction, resulting in large objects that are often impossible to store in computer memory and perform computation with. The bootstrap turns out to be a way to address this problem. Theory, simulations and a data example will be presented.

C0231: **Sparse estimation of function-on-function non-concurrent linear regression models**
Presenter: **Fabio Centofanti**, Universita di Napoli Federico II, Italy
Co-authors: Matteo Fontana, Antonio Lepore, Simone Vantini

Functional linear regression is the generalization of the classical regression analysis to the context of the functional data analysis. In particular, we focus on the function-on-function non-concurrent linear regression models (FonFLM) where both the regressor and the response variable have a functional form and where every point-wise evaluation of the response function possibly depends on every point-wise evaluation of the regressor function. No matter the sample size, FonFLMs are naturally over-parametrized and thus their estimation urges the introduction of penalization methods. We propose a new approach to the estimation of the regression operator of FonFLMs (i.e., S-LASSO) derived from the introduction in the loss function of both a LASSO like and a roughness penalization of the kernel of the regression operator that are able to annihilate the regression kernel on regions with smooth boundaries. After presenting some valuable asymptotic properties, we will then show by means of simulations that, with respect to other state-of-the-art techniques, the proposed estimator guarantees much higher interpretability, better predictive performances without any major loss in the estimation accuracy either the kernel is sparse or not.

C0267: **Investigating the temporal dynamics of tweets through functional data analysis**
Presenter: **Sara Fontanella**, Imperial College London, United Kingdom
Co-authors: Emiliano del Gobbo, Lara Fontanella, Luigi Ippoliti, Pasquale Valentini

In recent years, social media have become crucial tools for information dissemination. Microblogging online social network, such as Twitter, have provided a unique opportunity for people to communicate and share their interests and opinions on different topics. Their growth has attracted interest in many research fields and many analytical approaches to text mining have been developed to derive meaningful insights, valuable for the entire society. We propose an unsupervised classification approach rooted in functional data analysis aimed at analysing Twitter data streams. Our model aims at investigating the temporal dynamics of tweets, characterised by highly unstructured text, and, ultimately, identifying hidden patterns. We apply the proposed model to 10M tweets collected from January 2019 onwards to explore the engagement of active users in the Brexit debate and its temporal dynamics.

CO048 Room Business center FUNCTIONAL REGRESSION, DOMAIN SELECTION AND CLASSIFICATION Chair: Sara Sjostedt-de Luna

C0208: Statistical parametric mapping: One extra parameter to generalize classical hypothesis testing to continuum analysis

Presenter: **Todd Pataky**, Kyoto University, Japan

Gaussian random field theory (RFT) emerged in the statistics literature in the 1970s and by the early 1990s was widely adopted in neuroscience, where its applied form has become known as statistical parametric mapping (SPM). Although most widely used for the analysis of 4D brain image time series, SPM easily generalizes to univariate and multivariate continua of arbitrary dimensionality, and has hence been used in a variety of other fields, including biomechanics, morphometrics and cosmology. A brief historical context of SPM is provided, including the key RFT concepts underlying SPM, and subsequently developed nonparametric approximations to RFT distributions. Its capabilities for domain and feature selection will also be discussed. A few specific datasets will be explored, including interpretations of parametric SPM, nonparametric SPM, and functional data analysis. SPM's key strengths are that: (1) no continuum modeling is required, and (2) just a single smoothness parameter — in addition to the usual degrees of freedom — allows for parametric, continuum-level inference, and with negligible computational resources. SPM's key theoretical weakness is that it requires continuum registration robustness, only dealing with misregistration through sensitivity analysis and/or covariate modeling. Existing SPM software packages will be summarized along with implementation considerations.

C0229: Domain selection and family-wise error rate for functional data: A unified framework

Presenter: **Lina Schelin**, Umea University, Sweden

Co-authors: Konrad Abramowicz, Alessia Pini, Sara Sjostedt de Luna, Aymeric Stamm, Simone Vantini

Functional data are smooth, often continuous, random curves, which can be seen as an extreme case of multivariate data with infinite dimensionality. Just as component-wise inference for multivariate data naturally performs feature selection, subset-wise inference for functional data performs domain selection. We present a unified null-hypothesis testing framework for domain selection on populations of functional data. In detail, p-values of hypothesis tests performed on pointwise evaluations of functional data are suitably adjusted for providing a control of the family-wise error rate (FWER) over a family of subsets of the domain. Several state-of-the-art domain selection methods fit within this framework and differ from each other by the choice of the family over which the control is provided. In the existing literature, these families are always defined a priori. We also propose a novel approach, coined threshold-wise testing, where the family of subsets is built in a data-driven fashion. The method seamlessly generalizes to multidimensional domains in contrast to methods based on a-priori defined families. Theoretical results with respect to exactness, consistency, and strong and weak control of FWER for the methods within the unified framework are provided. The proposed local inferential techniques are applied to knee kinematic and brain tractography data.

C0241: Function-on-scalar regression for distributional responses

Presenter: **Alessandra Menafoglio**, Politecnico di Milano, Italy

Co-authors: Renata Talska, Karel Hron, Jitka Machalova, Eva Fiserova

The problem of performing functional linear regression is addressed when the response variable is represented as a probability density function (PDF). We consider the PDFs as functional compositional data, i.e., infinite-dimensional objects carrying relative information. We propose to embed the data in a Bayes Hilbert space, that allows to properly capture the key features of distributional data (e.g., scale invariance, relative scale). We develop a function-on-scalar regression model with distributional response. A centered log-ratio transformation is used to map the problem from the Bayes space to an unconstrained L2 space, where to develop computational methods based on B-splines for the model estimation. The methodological developments are shown on simulated and real data, dealing with metabolomics observations. Bootstrap methods are also discussed for the purpose of uncertainty quantification.

C0253: Functional data analysis-based sensitivity and domain-selective testing for integrated assessment models

Presenter: **Matteo Fontana**, Politecnico di Milano, Italy

Co-authors: Massimo Tavoni, Simone Vantini

Climate change is among the biggest threats to the survival of humankind. A lot of research efforts are being undertaken to study its evolution and the impact of policy measures aimed at mitigating its effects. The tool of choice for these tasks are Integrated Assessment Models (IAM), complex pieces of software that are used to predict socioeconomic variables for the next century. These models are black boxes de facto, due to the complex and nonlinear relationships between input and output variables. Moreover, they are often very computationally intensive, so it is virtually impossible to characterize the model response via standard Monte-Carlo based methods. To analyse models with these features, the most common choice is the use of data-parsimonious Global Sensitivity Analysis (GSA) techniques, that are moreover able to decouple additive and interaction effects. The main drawback of these techniques is that can deal only with univariate responses, disregarding the time dimension. The aim is to include the time dimension in the sensitivity analysis of IAMs, by modelling the time-variant outputs as smooth functions over the time domain, and then extend current GSA techniques to functional data. To assess the impact of model uncertainty on the significance of sensitivity indices, we frame the GSA as a Functional ANOVA problem, and then use a domain-selective permutational-based inferential technique for linear models to perform significance testing.

C0243: Efficient surface finish defect detection using reduced rank spline smoothers

Presenter: **Natalya Pya Arnqvist**, Umea University, Sweden

Co-authors: Blaise Ngendangenzwa, Leif Nilsson, Eric Lindahl, Jun Yu

One of the primary concerns of product quality control in the automotive industry is an automated detection of defects of small sizes on specular car body surfaces. A new statistical learning approach is presented for surface finish defect detection based on spline smoothing method for feature extraction and k-nearest neighbor probabilistic classifier. Rather than analyzing the natural images of the car body surfaces, the deflectometry technique is applied for image acquisition. Reduced rank cubic regression splines are used to smooth the pixel values while the effective degrees of freedom of the obtained smooths serve as components of the feature vector. A key advantage of the approach is that it allows us to reach near zero misclassification error when applying standard learning classifiers. We also propose the probability based performance evaluation metrics as alternatives to the conventional metrics. The usage of those provides the means for uncertainty estimation of the predictive performance of a classifier. Experimental classification results on the images obtained from the pilot system located at Volvo cab plant in Umea, Sweden, show that the proposed approach is much more efficient than compared methods.

CO052 Room Triton 1+2 INFERENCE FOR DENSITIES, MODES, AND STRUCTURES Chair: Jochen Einbeck

C0230: Resampling under unimodality via the empirical zero bias transform

Presenter: **Bruno Ebner**, Karlsruhe Institute of Technology, Germany

A new strictly nonparametric method for resampling is proposed under the assumption of unimodality. The method is based on the so called empirical zero bias transform. We show that it improves the finite sample power of classical tests like the Silverman test as well as the DIP test under alternatives while exploiting the significance level under the null hypothesis. We show the efficiency in a Monte Carlo simulation and that it is a serious competitor to the only existing parametric method by Cheng and Hall.

C0168: **Goodness-of-fit test for bivariate models of time series of counts**

Presenter: **Sarka Hudecova**, Charles University, Prague, Czech Republic

Several types of bivariate (multivariate) models for integer-valued time series have been suggested recently in the literature. These involve both INAR type models (based on the thinning operator) as well as INGARCH type models (based on specification of the conditional distribution). We propose a goodness of fit test for two important classes of bivariate models: INAR(1) model and Poisson INARCH(1) model constructed via a trivariate reduction method. The proposed test is based on the probability generating function as this has already proved to be a useful tool in testing goodness-of-fit and detecting a change point in univariate integer-valued time series.

C0166: **The Christmas tree plot: A graphical tool for assessing the suitability of a count regression model**

Presenter: **Paul Wilson**, University Of Wolverhampton, United Kingdom

Co-authors: Jochen Einbeck

Whilst many numeric methods, such as AIC and deviance, exist for assessing or comparing model fit, diagrammatic methods are few. We present a diagnostic plot, which we refer to as a 'Christmas tree plot' due its characteristic shape, that may be used to visually assess the suitability of a given count data model. The graphical tool presented is an alternative approach to assessing model fit, which may be used on its own, or in conjunction with other methods. It enables the user to determine whether the observed frequency of a given count in the data is compatible with that to be expected under a given distribution, and thus, if non-compatibility of the model with the data is indicated, the nature of the incompatibility is apparent, possibility leading to the determination of a more suitable model.

C0190: **Some thoughts on the estimation of antimodes**

Presenter: **Jochen Einbeck**, Durham University, United Kingdom

While there is some considerable literature on the estimation of modes of a continuous density function from sample data, the problem of estimating the antimodes (that is, local minima of the density between the modes) has received very little attention. Antimodes can be considered as points, within the support of the density, which are less likely to be observed than any point around them. Thus, they are of obvious relevance for questions such as resource allocation, and furthermore they can be interpreted as decision or classification boundaries, if each local mode is associated with a class. We discuss some possible approaches for the estimation of antimodes, starting with the univariate case. If the density can be represented by a finite Gaussian mixture, approximate expressions for the antimode can be obtained, which are formally equivalent to Bayesian Quadratic Discriminant Analysis. In the more general case, for fully nonparametric densities, antimodes can be estimated by an inverse version of the mean shift algorithm, notably without requiring the estimation of the density itself.

C0270: **Infinite mixtures of beta regression models for bounded-domain variables**

Presenter: **Ioannis Kosmidis**, University of Warwick and The Alan Turing Institute, United Kingdom

Co-authors: Achim Zeileis

Beta regression is a useful tool for modelling bounded-domain continuous response variables, such as proportions, rates fractions and concentration indices. One important limitation of beta regression models is that they do not apply when at least one of the observed responses is on the boundary — in such scenarios the likelihood function is simply 0 regardless of the value of the parameters. The relevant approaches in the literature focus on either the transformation of the observations by small constants so that the transformed responses end up in the support of the beta distribution, or the use of a discrete-continuous mixture of a beta distribution and point masses at either or both of the boundaries. The former approach suffers from the arbitrariness of choosing the additive adjustment. The latter approach gives a "special" interpretation to the boundary observations relative to the non-boundary ones, and requires the specification of an appropriate regression structure for the hurdle part of the overall model, generally leading to complicated models. We rethink of the problem and present an alternative model class that leverages the flexibility of the beta distribution, can naturally accommodate boundary observations and preserves the parsimony of beta regression, which is a limiting case. Likelihood-based estimation and inferential procedures for the new model are presented, and its usefulness is illustrated by applications.

Sunday 14.04.2019 14:20 - 15:50 Parallel Session C – CRONOSMDA2019

CI015 Room Triton 3 SPRING COURSE SESSION II Chair: Karel Hron

C0167: **Applied compositional data analysis**
Presenter: **Karel Hron**, Palacky University, Czech Republic

Compositional data are multivariate observations that carry relative information. They are measured in units like proportions, percentages, mg/l, mg/kg, ppm, and so on, i.e., as data that might obey (or not) a constant sum of components. Due to their specific features, the statistical analysis of compositional data must obey the geometry of the simplex sample space. In order to enable processing of compositional data using standard statistical methods, compositions can be conveniently expressed by means of real vectors of logratio coordinates. Their meaningful interpretability is of primary importance in practice. Aim of the course is to introduce the logratio methodology of compositional data together with a wide range of its possible applications. The first part of the course will be devoted to theoretical aspects of the methodology including principles of compositional data analysis, geometrical representation of compositions, construction of logratio coordinates and their interpretability. In the second part exploratory data analysis including visualization will be presented, followed by concrete popular statistical methods, e.g. correlation and regression analysis, or principal component analysis, and even methods for processing of high-dimensional data adapted within the logratio methodology. Also robust counterparts to some of these methods will be discussed. Numerical examples will be presented using the package robCompositions of the statistical software R.

CO027 Room Neptune FUNCTIONAL DATA AND DEPTHS Chair: Gil Gonzalez-Rodriguez

C0172: **Nonparametric covariance estimation for mixed longitudinal studies**
Presenter: **Kehui Chen**, University of Pittsburgh, United States
Co-authors: Anru Zhang

Motivated by applications of mixed longitudinal studies, where a group of subjects entering the study at different ages (cross-sectional) are followed for successive years (longitudinal), we consider nonparametric covariance estimation with samples of noisy and partially-observed functional trajectories. We will introduce a novel sequential aggregation scheme, which works for both dense regular and sparse irregular observations. We will present numerical experiment results and applications a midlife women's working memory study. We will also discuss the details of identifiability and estimation consistency.

C0188: **The halfspace depth characterization problem**
Presenter: **Stanislav Nagy**, Charles University, Czech Republic

The halfspace depth is an inferential tool that aims to generalize quantiles to multivariate datasets. It has been long conjectured that, just as for the usual quantiles, there is a one-to-one relation between all Borel probability measures, and all possible depth surfaces. We answer this conjecture in the negative. That suggests an interesting open problem of characterizing those probability measures that possess a unique depth. A complete solution to this problem would have far-reaching implications, not only in the theory of multivariate statistics.

C0215: **Functional change point detection for fMRI data**
Presenter: **Claudia Kirch**, Otto-von-Guericke University Magdeburg, Germany
Co-authors: John Aston, Christina Stoehr

Functional magnetic resonance imaging (fMRI) is now a well-established technique for studying the brain. However, in many situations, such as when data are acquired in a resting state, the statistical analyzes depends crucially on stationarity which could easily be violated. We introduce tests for the detection of deviations from this assumption by making use of change point alternatives, where changes in the mean as well as covariance structure of functional time series are considered. Because of the very high-dimensionality of the data an approach based on a general covariance structure is not feasible, such that computations will be conducted by making use of a multidimensional separable functional covariance structure. Using the developed methods, a large study of resting state fMRI data is conducted to determine whether the subjects undertaking the resting scan have nonstationarities present in their time courses. It is found that a sizeable proportion of the subjects studied are not stationary.

C0284: **Spatial regression with PDE penalization: Consistency of the estimator**
Presenter: **Eleonora Arnone**, Politecnico di Milano, Italy
Co-authors: Alois Kneip, Fabio Nobile, Laura Sangalli

The consistency of the estimator in Spatial Regression with Partial Differential Equation penalization method (SR-PDE) is studied. SR-PDE is a technique for the estimation of a spatial dependent field over a two-dimensional complex domain from pointwise noisy observations when prior information on the field is available in form of a PDE. The consistency is studied both for the estimator in the infinite dimensional setting and for the discrete estimator obtained with finite elements method. Bias and variance of the estimator are analyzed with respect to the sample size and the value of the smoothing parameter. It is shown that optimal rates of convergence can be reached for the mean squared error in the L^2 and discrete norm when the number of observations goes to infinity. Simulation studies to verify the convergence rates are performed in a simple setting.

CO046 Room Business center EXTREMES FOR MULTIVARIATE DATA Chair: Gilles Stupfler

C0165: **Hypothesis testing for tail dependence parameters on the boundary of the parameter space**
Presenter: **Anna Kiriliouk**, University of Namur, Belgium

Modelling multivariate tail dependence is one of the key challenges in extreme-value theory. Multivariate extremes are usually characterized using parametric models, some of which have simpler submodels at the boundary of their parameter space. We propose hypothesis tests for tail dependence parameters that, under the null hypothesis, are on the boundary of the alternative hypothesis. The asymptotic distribution of the weighted least squares estimator is given when the true parameter vector is on the boundary of the parameter space, and a deviance- and Wald-type test statistic are proposed. The performance of these test statistics is evaluated for the Brown-Resnick model and the max-linear model. In particular, simulations show that it is possible to recover the number of factors used in a max-linear model. Finally, the methods are applied to characterize the dependence structure of two major stock market indices, the DAX and the CAC40.

C0206: **On a relationship between randomly and non-randomly thresholded empirical average excesses for heavy tails**
Presenter: **Gilles Stupfler**, The University of Nottingham, United Kingdom

Motivated by theoretical similarities between the classical Hill estimator of the tail index of a heavy-tailed distribution and one of its pseudo-estimator versions featuring a non-random threshold, we show a novel asymptotic representation of a class of empirical average excesses above a high random threshold, expressed in terms of order statistics, using their counterparts based on a suitable non-random threshold, which are sums of independent and identically distributed random variables. As a consequence, the analysis of the joint convergence of such empirical average excesses essentially boils down to a combination of Lyapunov's central limit theorem and the Cramer-Wold device. We illustrate how this allows us to improve upon, as well as produce conceptually simpler proofs of, very recent results about the joint convergence of marginal Hill estimators for a random vector with heavy-tailed marginal distributions. These results are then applied to the proof of a convergence result for a tail index estimator when the heavy-tailed variable of interest is randomly right-truncated. New results on the joint convergence of conditional tail moment estimators of a random vector with heavy-tailed marginal distributions are also obtained.

C0209: **Estimation of conditional extreme risk measures from heavy-tailed elliptical random vectors**
Presenter: **Antoine Usseglio-Carleve**, Inria, France

The focus is on some conditional extreme risk measures estimation for elliptical random vectors. Previously, we had proposed a methodology to approximate extreme quantiles, based on two extremal parameters. We thus propose some estimators for these parameters, and study their consistency and asymptotic normality in the case of heavy-tailed distributions. Thereafter, from these parameters, we construct extreme conditional quantiles estimators, and give some conditions that ensure consistency and asymptotic normality. Using recent results on the asymptotic relationship between quantiles and other risk measures, we deduce estimators for extreme conditional Lp-quantiles and Haezendonck-Goovaerts risk measures. Under similar conditions, consistency and asymptotic normality are provided. In order to test the effectiveness of our estimators, we propose a simulation study. A financial data example is also proposed.

C0210: **Cluster-based extremal inference for multivariate time series**
Presenter: **Anja Janssen**, KTH Royal Institue of Technology, Sweden
Co-authors: Holger Drees

Statistical procedures for inference on extremal properties of a multivariate time series are affected by the underlying extremal dependence structures. Many common time series models exhibit a clustering of extreme values and this will typically affect the variance of estimators which were built for i.i.d. observations. On the other hand, the behavior of quantities of interest, for example marginal distributions of the spectral tail process,is closely related to the overall dependence structure which we see in extremal clusters. We explore how this connection can be exploited to derive new estimators for extremal quantities.

CO039 Room Triton 1+2 DEEP LEARNING AND BAYESIAN ANALYSIS Chair: Florian Frommlet

C0161: **Bayesian deep learning**
Presenter: **Pietro Michiardi**, EURECOM, France
Co-authors: Maurizio Filippone

Drawing meaningful conclusions on the way complex real life phenomena work and being able to predict the behavior of systems of interest requires developing accurate and highly interpretable mathematical models whose parameters need to be estimated from observations. In modern applications of data modeling, however, we are often challenged with the lack of such models, and even when these are available they are too computational demanding to be suitable for standard parameter optimization/inference. Deep learning techniques have become extremely popular to tackle such challenges in an effective way, but they do not offer satisfactory performance in applications where quantification of uncertainty is of primary interest. Bayesian Deep Learning techniques have been proposed to combine the representational power of deep learning techniques with the ability to accurately quantify uncertainty thanks to their probabilistic treatment. While attractive from a theoretical standpoint, the application of Bayesian Deep Learning techniques poses huge computational challenges that arguably hinder their wide adoption. New trends in Bayesian Deep Learning will be presented, with particular emphasis on practical and scalable inference techniques and applications.

C0162: **A novel approach to deep Bayesian regression**
Presenter: **Florian Frommlet**, Medical University Vienna, Austria
Co-authors: Aliaksandr Hubin, Geir Olve Storvik

One of the most exciting recent developments in data analysis is deep learning. Multilayer networks have become extremely successful in performing prediction tasks and are successfully applied in many different areas. However, the resulting prediction models are often difficult to interpret and potentially suffer from overfitting. The aim is to bring the ideas of deep learning into a statistical framework which yields more parsimonious models and allows us to quantify model uncertainty. To this end, we introduce the class of deep Bayesian regression models (DBRM) consisting of a generalized linear model combined with a comprehensive non-linear feature space, where non-linear features are generated just like in deep learning. DBRM can easily be extended to include latent Gaussian variables to model complex correlation structures between observations, which seems to be not easily possible with existing deep learning approaches. Two different algorithms based on MCMC are introduced to fit DBRM and to perform Bayesian inference. The predictive performance of these algorithms is compared with a large number of state of the art learning algorithms. Furthermore we illustrate how DBRM can be used for model inference in various applications.

C0163: **Combining model and parameter uncertainty in Bayesian neural networks**
Presenter: **Aliaksandr Hubin**, Norwegian Computing Center, Norway

Bayesian Neural Networks (BNNs) have recently regained a significant amount of attention in the deep learning community due to the development of scalable approximate Bayesian inference techniques for training them. The advantage of using BNNs is straightforward: parameter and prediction uncertainty become easily available allowing to perform rigid statistical analysis. However, so far there have been no scalable techniques capable of combining both model and parameter uncertainty developed. We introduce the concept of model uncertainty in Bayesian neural networks and hence make inference in the joint space of models and parameters. Furthermore, we suggest adaptation of a scalable variational inference approach with reparametrizations of marginal inclusion probabilities to incorporate the model space constraints.

C0184: **Deep structured prediction**
Presenter: **Anton Osokin**, National Research University Higher School of Economics, Russia

Two approaches are discussed to use neural networks for making joint predictions of many variables. First, we will touch combining neural networks with classical techniques such as Conditional Random Fields (CRFs). We will cover using optimization algorithms as structured pooling, unrolling of algorithm iterations into network layers and direct differentiation of the output w.r.t. the input. Secondly, we well discuss task specific ways of training deep auto-regressive models (such as seq2seq). We will illustrate all the discussed methods with applications from computer vision and natural language processing.

Sunday 14.04.2019 16:20 - 17:30 Parallel Session D – CRONOSMDA2019

CI007 Room Triton 1+2 CRoNoS Session I Chair: Ori Davidov

C0175: **Robust methods for fuzzy clustering of non-linear structures**
Presenter: **Maria Brigida Ferraro**, Sapienza University of Rome, Italy
Co-authors: Paolo Giordani

In many practical situations data may be characterized by non-linear structures. Classical (hard or fuzzy) algorithms usually detect clusters by computing the Euclidean distance among pairs of objects. They are based on the linearity assumption and, therefore, do not identify properly clusters characterized by non-linear structures. In order to overcome this limitation, the so-called geodesic distance, able to capture and preserve the intrinsic geometry of the data, can be considered. A new fuzzy relational clustering algorithm based on the geodesic distance is introduced. In addition, to improve its adequacy, a robust version is proposed.

C0193: **Distances for clustering non-precise information: A comparative study**
Presenter: **Ana Belen Ramos-Guajardo**, University of Oviedo, Spain
Co-authors: Maria Brigida Ferraro

Different clustering methods for non-precise information have been developed in the recent decades. Some of those methods include also fuzziness in the process. This is the case of the well-known fuzzy k-means procedure for clustering fuzzy numbers. The distance between fuzzy numbers employed is basically defined as a weighted sum of the squared Euclidean distances between their mid-points and their spreads. Nevertheless, the fuzzy k-means approach does not allow for the correlation structure between variables, which is a shortcoming whenever the shape of the clusters is not spherical. For this reason, the Mahalanobis distance involving the corresponding covariance matrices between the variables has been introduced, and a fuzzy clustering approach based on that distance is proposed. Both methodologies are compared by means of simulation studies, and a real-life situation is also tackled.

C0176: **Approximations for extremes and reliability of high-dimension coherent systems**
Presenter: **Ivette Gomes**, FCiencias.ID, Universidade de Lisboa and CEAUL, Portugal
Co-authors: Sonia Dias, Luisa Canto e Castro, Paula Reis

The rate of convergence of linearly normalized maxima/minima to the corresponding non-degenerate extreme value (EV) limiting distribution, either univariate or multivariate, is a relevant problem in the field of extreme value theory. Moreover, it is well known that when dealing with univariate extremes, the limiting EV approximation can be asymptotically improved, through the so-called penultimate approximations, which have been theoretically studied from different perspectives. But despite of not yet fully developed, similar penultimate approximations obviously appear in the multivariate case. The aforementioned approximations will be revisited in the field of reliability, where any coherent system can be represented as either a series-parallel (SP) or a parallel-series system (PS), with a lifetime that can thus be written as the minimum of maxima or the maximum of minima. The identification of the possible limit laws for the system reliability of homogeneous and non-homogeneous PS systems is sketched and the gain in accuracy is assessed whenever a penultimate approximation is used instead of the ultimate limiting one.

CI017 Room Triton 3 Spring course session III Chair: Tim Verdonck

C0259: **Robust high-dimensional data analysis**
Presenter: **Tim Verdonck**, KU Leuven, Belgium
Co-authors: Stefan Van Aelst

Robust statistics develops methods and techniques to reliably analyze data in the presence of outlying measurements. Next to robust inference outlier detection is also an important goal of robust statistics. When analyzing high-dimensional data sparse solutions are often desired to enhance interpretability of the results. Moreover, when the data are of uneven quality robust estimators are needed that are computationally efficient such that solutions can be obtained in a reasonable amount of time. Moreover, if many variables in high-dimensional data can have some anomalies in their measurements, then it is not reasonable anymore to assume that a majority of the cases is completely free of contamination. In such cases the standard paradigm of robust statistics is not valid anymore, but alternative methods need to be used. Robust procedures for high-dimensional data, such as estimation of location and scatter, linear regression, generalized linear models and principal component analysis. The good performance of these methods is illustrated on real data using R.

CO025 Room Neptune Functional/high-dimensional statistics I Chair: Frederic Ferraty

C0257: **Joint sparse clustering and alignment of functional data: Theory and case studies**
Presenter: **Valeria Vitelli**, University of Oslo, Norway

Finding sparse solutions to clustering problems has emerged as a hot topic in statistics in recent years, and it very recently emerged in the functional data literature, too: it is often of much interest to cluster the curves while jointly detecting their most relevant features. Functional sparse clustering is analytically defined as a variational problem, where a constraint ensures the sparsity of the solution. This problem is well-posed and has a unique optimal solution. When dealing with curve clustering, misalignment is a frequent problem: in such case, the only possible approach is to first align the curves, and then use a sparse functional clustering method to estimate the groups and select the domain. However, it is well-known that aligning and clustering the curves jointly is beneficial for the analysis, and many methods to jointly cluster and align curves have already been proposed in the literature on functional data. By focusing on one of these methods, we propose a possible approach to jointly deal with sparse functional clustering while also aligning the curves. We thus propose a novel algorithm which jointly performs all these tasks: clustering, alignment, and domain selection. We prove the well-posedness of the problem, and test the method on simulated data. We also show some results of the analysis with well-known benchmark functional datasets. We conclude with a vision of the possible future research directions on the topic.

C0256: **Categorical functional data analysis with R**
Presenter: **Cristian Preda**, University of Lille, France

Categorical functional data represented by paths of a stochastic jump process with continuous time are considered for dimension reduction (visualisation), regression and clustering. A simulation study and an analysis of discharge medical letters are presented in an R framework.

C0251: **Stability of a network inference procedure in high-dimension**
Presenter: **Emilie Devijver**, CNRS, France
Co-authors: Melina Gallopin, Remi Molinier

Network inference is widely utilized to evaluate and represent dependencies between continuous variables. Gaussian graphical models have been developed the last years, tackling the high-dimension problem through several assumptions. The focus is on the stability of a procedure called shock, which infers a modular network represented by a block-diagonal covariance matrix. This structure has strong advantages, among such reducing the dimension, facilitating the interpretation and being stable. The stability of the procedure is supported by strong theoretical guarantees based on topological tools, intensive simulations and real data analysis.

CP001 Room Business center POSTER SESSION **Chair: Elena Fernandez Iglesias**

C0258: **Simulation of multidimensional stochastic integrals driven by Levy motions**
Presenter: **Dmitry Otryakhin**, Aarhus University, Denmark

Stochastic integrals driven by Levy motions and possessing stationarity property have been used to model turbulence in recent years. Two simulation methods, based on previous ones, are developed for that type of processes. The algorithms are compared in terms of applicability, speed and numerical error.

C0288: **Estimating pediatric hypertension prevalence in Portugal**
Presenter: **Maria Filomena Teodoro**, CINAV - Portuguese Naval Academy/CEMAT, Instituto Superior Tecnico, Portugal

Pediatric hypertension may silently appear in childhood and its diagnostic is based on regular blood pressure registers, that usually can depend on age, gender and weight. This disease can evolve and usually is associated with severe organ-damage, being important to be aware about its existence to conduce an early intervention. To evaluate the prevalence of the disease in Portugal, it was applied a questionnaire to a representative sample of portuguese population. The data collection took a long period of time to be concluded and resulted from the commitment of numerous professionals involved in health, namely doctors, nurses, trainees, etc. The data was statistically analyzed and modeled, in order to infer about the existence of possible associations between socio-demographic variables (age, gender, race, residence, rea, caregivers graduation level and occupation) and the disease occurrence. Some statistically significant associations between socio-demographic variables and high blood pressure occurrence were achieved. We have applied generalized linear models and other multivariate, adequate methods to model and analyze the kind of data under study.

Sunday 14.04.2019 17:40 - 18:50 Parallel Session E – CRONOSMDA2019

CI019 Room Triton 3 SPRING COURSE SESSION IV Chair: Tim Verdonck

C0183: **Robust high-dimensional data analysis**
Presenter: **Tim Verdonck**, KU Leuven, Belgium
Co-authors: Stefan Van Aelst

Robust statistics develops methods and techniques to reliably analyze data in the presence of outlying measurements. Next to robust inference outlier detection is also an important goal of robust statistics. When analyzing high-dimensional data sparse solutions are often desired to enhance interpretability of the results. Moreover, when the data are of uneven quality robust estimators are needed that are computationally efficient such that solutions can be obtained in a reasonable amount of time. Moreover, if many variables in high-dimensional data can have some anomalies in their measurements, then it is not reasonable anymore to assume that a majority of the cases is completely free of contamination. In such cases the standard paradigm of robust statistics is not valid anymore, but alternative methods need to be used. Robust procedures for high-dimensional data, such as estimation of location and scatter, linear regression, generalized linear models and principal component analysis. The good performance of these methods is illustrated on real data using R.

CO094 Room Neptune FUNCTIONAL/HIGH-DIMENSIONAL STATISTICS II Chair: Frederic Ferraty

C0246: **Functional linear spatial autoregressive modeling with endogenous weight matrix**
Presenter: **Sophie Dabo**, University of Lille, France
Co-authors: Zied Gharbi

A functional linear autoregressive spatial model with endogenous weight matrix is proposed where the explanatory variable takes values in a functional space and the response process is real valued and spatially autocorrelated. The particularity of the model is that the spatial correlation defined by a weight matrix depends not only on the geographic distances between neighbours but also on others factors such as economic distances. So the weight matrix is no more exogenous as supposed in conventional spatial autoregressive model (SAR). We introduce a two steps estimation method that consists in reducing the infinite dimension of the predictor and applying a quasi-maximum likelihood estimator. We establish consistency and asymptotic normality of the regression functional parameter estimate and illustrate the skills of the methods on simulated data as well as on application to real data.

C0247: **A general white noise test based on kernel lag-window estimates of the spectral density operator**
Presenter: **Vaidotas Characiejus**, Universite libre de Bruxelles, Belgium
Co-authors: Gregory Rice

A general white noise test for functional time series is considered. The idea of the test is to estimate a distance between the spectral density operator of a weakly stationary time series and the constant spectral density operator of an uncorrelated time series. The estimator of the distance is based on a kernel lag-window type estimator of the spectral density operator. When the observed time series is a strong white noise in a real separable Hilbert space, it is shown that the asymptotic distribution of the test statistic is standard normal, and it is further shown that the test statistic diverges for general serially correlated time series. These results recover as special cases some previous tests. In order to implement the test, a number of kernel and bandwidth choices is proposed and studied, including a new data adaptive bandwidth, as well as data adaptive power transformations of the test statistic that improve the normal approximation in finite samples. A simulation study demonstrated that the proposed method has good size and improved power when compared to other methods available in the literature, while also offering a light computational burden. The utility of the proposed test is demonstrated by considering an application to daily Eurodollar futures curves.

C0197: **Estimation of extreme expectiles given a high-dimensional covariate**
Presenter: **Stephane Girard**, Inria, France
Co-authors: Gilles Stupfler

Expectiles are least-square analogues of quantiles. They have received a fair amount of attention due to their potential for application in financial, actuarial, and economic contexts. Some recent work has focused on the application of extreme expectiles to assess tail risk, and on their estimation in a heavy-tailed framework. We investigate the estimation of extreme expectiles of a heavy-tailed random variable Y given a high-dimensional covariate X. We derive generic conditions under which the limiting behaviour of our estimators can be established. Applications are presented to some regression models. A finite-sample study illustrates the behaviour of our procedures in practice.

CO080 Room Business center MODEL-BASED AND MULTIVARIATE FUNCTIONAL DATA Chair: Cristian Preda

C0291: **Gaussian-based visualization of Gaussian and non-Gaussian model-based clustering**
Presenter: **Vincent Vandewalle**, Inria, France
Co-authors: Christophe Biernacki, Matthieu Marbac

A generic method is introduced to visualize in a Gaussian-like way, and onto Rd, results of Gaussian or non-Gaussian model-based clustering. The key point is to explicitly force a spherical Gaussian mixture visualization to inherit from the within cluster overlap which is present in the initial clustering mixture. The result is a particularly user-friendly draw of the clusters, allowing any practitioner to have a thorough overview of the potentially complex clustering result. An entropic measure allows us to inform of the quality of the drawn overlap, in comparison to the true one in the initial space. The proposed method is illustrated on four real data sets of different types (categorical, mixed, functional and network) and is implemented on the R package ClusVis.

C0242: **Functional mixtures-of-experts**
Presenter: **Faicel Chamroukhi**, Caen University, Lab of Mathematics LMNO, France

Mixtures-of-Experts (MoE) is a family of mixture models which are effective for modeling heterogeneous data in regression, classification and clustering. They model complex non-linear relationships between a response Y and a predictor X in situations where they arise from a heterogeneous population of K homogeneous sub-populations governed by a latent structure represented by an unknown variable Z. They have been widely studied in the multivariate analysis literature where the input or the response are possibly multi-dimensional vectors. We present a new framework of Mixtures-of-Experts models in which the input $X(t)$ and/or the output $Y(t)$ can be functions (i.e, curves) possibly multivariate. The proposed functional Mixtures-of-Experts (FunME) models allow to capture the nonlinear relationship between a set of input and output functional variables which are observed continuously (i.e, over time for time series) and thus which are of infinite dimension. We propose a maximum likelihood fitting based on the EM algorithm.

C0202: **A functional model-adjusted spatial scan statistic**
Presenter: **Michael Genin**, University of Lille, France
Co-authors: Mohamed Salem Ahmed

A functional model-adjusted spatial scan statistic is introduced to adjust the detection of clusters on longitudinal confounding factors indexed in space. An approach based on generalized linear functional models is used to construct this spatial scan statistic, with longitudinal confounding factors being considered as functional covariates. A general framework is proposed for various probability models and application to the Poisson model shows that this method is equivalent to a classical statistical spatial scan statistic considering an underlying population adjusted for covariates. Through a simulation study, we show that this method has a better quality of adjustment than other methods based on univariate and multivariate models. The proposed method is illustrated using premature mortality data in France during the period from 1998 to 2013, considering the quarterly unemployment rate as a longitudinal confounding factor.

CO058 **Room Triton 1+2** **BAYESIAN ANALYSIS OF COMPLEX DATA** **Chair: Bernardo Nipoti**

C0213: **Varying-sparsity regression models with application to cancer proteogenomics**
Presenter: **Francesco Stingo**, University of Florence, Italy

Identifying patient-specific prognostic biomarkers is of critical importance in developing personalized treatment for clinically and molecularly heterogeneous diseases such as cancer. We propose a novel regression framework, Bayesian hierarchical varying-sparsity regression (BEHAVIOR) models to select clinically relevant disease markers by integrating proteogenomic (proteomic+genomic) and clinical data. Our methods allow flexible modeling of proteingene relationships as well as induces sparsity in both protein-gene and protein-survival relationships, to select genomically driven prognostic protein markers at the patient-level. Simulation studies demonstrate the superior performance of BEHAVIOR against competingmethod in terms of both protein marker selection and survival prediction. We apply BEHAVIOR to The Cancer Genome Atlas (TCGA) proteogenomic pan-cancer data and find several interesting prognostic proteins and pathways that are shared across multiple cancers and some that exclusively pertain to specific cancers.

C0232: **Colombian women's life choices: A Bayesian nonparametric multivariate regression approach**
Presenter: **Isadora Antoniano-Villalobos**, Ca' Foscari University of Venice, Italy
Co-authors: Andrea Cremaschi, Raffaella Piccarreta, Sara Wade

Women in the Latin America and Caribbean countries face difficulties related to the patriarchal traits of their society. In Colombia, the well-known conflict afflicting the country since 1948, has increased the risk of vulnerable groups. It is important to determine if recent efforts to improve the welfare of women have had a positive effect extending beyond the capital, Bogota. In an initial effort to shed life on this matter, we analyze cross-sectional data arising from the Demographic and Health Survey Program which collects and disseminates data on random samples of households selected from a national sampling frame. Our aim is to study the relationship between baseline socio-demographic factors and variables associated to fertility, partnership patterns and work activity. We propose a flexible Bayesian nonparametric multivariate regression model, which can capture nonlinear regression functions and the presence of non-normal errors, such as heavy tails or multi-modality. The model has interpretable covariate-dependent weights constructed through normalization, allowing for combinations of both categorical and continuous covariates, as well as censoring in one or more of the responses. Computational difficulties for inference are overcome through an adaptive truncation algorithm combining adaptive Metropolis-Hastings and sequential Monte Carlo to create a sequence of automatically truncated posterior mixtures.

C0249: **Distributional properties of Bayesian deep neural networks**
Presenter: **Julyan Arbel**, Inria, France

Deep Bayesian neural networks with Gaussian priors on the weights and ReLU-like nonlinearities are investigated, shedding light on novel regularization mechanisms at the level of the units of the network, both pre- and post-nonlinearities. The main thrust is to establish that the units prior distribution becomes increasingly heavy-tailed with depth. We show that first layer units are Gaussian, second layer units are sub-Exponential, and we introduce sub-Weibull distributions to characterize the deeper layers units. Bayesian neural networks with Gaussian priors are well known to induce the weight decay penalty on the weights. In contrast, our result indicates a more elaborate regularization scheme at the level of the units. This result provides new theoretical insight on deep Bayesian neural networks, underpinning their natural shrinkage properties and practical potential.

Sunday 14.04.2019 18:50 - 19:20 Parallel Session F – CRONOSMDA2019

CI084 Room Triton 3 SPRING COURSE SESSION V Chair: Stefan Van Aelst

C0205: **Robust high-dimensional data analysis**

Presenter: **Stefan Van Aelst**, University of Leuven, Belgium

Co-authors: Tim Verdonck

Robust statistics develops methods and techniques to reliably analyze data in the presence of outlying measurements. Next to robust inference outlier detection is also an important goal of robust statistics. When analyzing high-dimensional data sparse solutions are often desired to enhance interpretability of the results. Moreover, when the data are of uneven quality robust estimators are needed that are computationally efficient such that solutions can be obtained in a reasonable amount of time. Moreover, if many variables in high-dimensional data can have some anomalies in their measurements, then it is not reasonable anymore to assume that a majority of the cases is completely free of contamination. In such cases the standard paradigm of robust statistics is not valid anymore, but alternative methods need to be used. Robust procedures for high-dimensional data, such as estimation of location and scatter, linear regression, generalized linear models and principal component analysis. The good performance of these methods is illustrated on real data using R.

Monday 15.04.2019 08:50 - 10:00 Parallel Session G – CRONOSMDA2019

CI086 Room Triton 3 SPRING COURSE SESSION VI Chair: Stefan Van Aelst

C0265: **Robust high-dimensional data analysis**
Presenter: **Stefan Van Aelst**, University of Leuven, Belgium
Co-authors: Tim Verdonck

Robust statistics develops methods and techniques to reliably analyze data in the presence of outlying measurements. Next to robust inference outlier detection is also an important goal of robust statistics. When analyzing high-dimensional data sparse solutions are often desired to enhance interpretability of the results. Moreover, when the data are of uneven quality robust estimators are needed that are computationally efficient such that solutions can be obtained in a reasonable amount of time. Moreover, if many variables in high-dimensional data can have some anomalies in their measurements, then it is not reasonable anymore to assume that a majority of the cases is completely free of contamination. In such cases the standard paradigm of robust statistics is not valid anymore, but alternative methods need to be used. Robust procedures for high-dimensional data, such as estimation of location and scatter, linear regression, generalized linear models and principal component analysis. The good performance of these methods is illustrated on real data using R.

CO035 Room Neptune ADVANCES IN FUNCTIONAL DATA ANALYSIS Chair: Dominik Liebl

C0239: **Functional principal component analysis for Lorenz curves**
Presenter: **Enea Bongiorno**, Universita del Piemonte Orientale, Italy
Co-authors: Aldo Goia

Lorenz curves are widely used in economic studies (inequality, poverty, differentiation, etc.). From a model point of view, such curves can be seen as constrained functional data for which functional principal component analysis (FPCA) could be defined. Although statistically consistent, performing FPCA using the original data can lead to a suboptimal analysis from a mathematical and interpretation point of view. In fact, the family of Lorenz curves lacks very basic (e.g., vectorial) structures and, hence, must be treated with ad hoc methods. The aim is to provide a rigorous mathematical framework via an embedding approach to define a coherent FPCA for Lorenz curves. This approach is used to explore a functional dataset from the Bank of Italy income survey.

C0248: **Penalized smoothing for density estimation and its implications for risk management**
Presenter: **Michelle Carey**, Univerity College Dublin, Ireland
Co-authors: James Ramsay, Christian Genest

Pearson's system of curves is a flexible class of continuous univariate distributions which includes many classical models. It can accommodate most combinations of mean, variance, skewness, and kurtosis. Each density f in this class is the unique solution to the differential equation (ODE) that depends on the choice of the parameters of the ODE. Estimating f from random observations is a challenging problem and the current approaches often fail to produce meaningful estimators. It is shown that both the parameters of the ODE and f can be estimated through an adaptation of a model-based smoothing procedure that incorporates differential equations. The resulting estimate of f is the distribution within Pearson's wide class that best represents the data. The approach is illustrated using data on the TSX composite index from the 2008 financial crisis. Estimates of the Value-at-Risk and Expected Shortfall based on f are shown to outperform the estimates currently used by financial institutions and regulators for market risk assessment.

C0262: **A test of equality of distributions for Hilbert-valued random elements**
Presenter: **Gil Gonzalez-Rodriguez**, University of Oviedo, Spain
Co-authors: Ana Colubi, Wenceslao Gonzalez-Manteiga

Two independent random elements taking on values in a separable Hilbert space are considered. The aim is to develop a bootstrap test to check whether they have the same distribution or not. A common transformation of both random elements into a new separable Hilbert space is considered in such a way that the equality of expectations of the transformed random elements is equivalent to the equality of distributions. Thus, a bootstrap procedure to check the equality of means can be used in order to solve the original problem. It will be shown that the test can be solved with simple operations in the original space, without the need for applying the mentioned transformation.

CO062 Room Triton 1+2 HIGH DIMENSIONAL METHODS FOR COMPLEX DATA Chair: Ivor Cribben

C0185: **Modeling evolution of spectral properties in stationary processes of varying dimensions**
Presenter: **Raanju Sundararajan**, KAUST, Saudi Arabia
Co-authors: Hernando Ombao

Analysis of multivariate time series, stationary and nonstationary, often involves a linear decomposition of the observed series into latent sources. Methods like PCA, ICA and Stationary Subspace Analysis (SSA) assume the observed multivariate process is linearly generated by latent sources that can be stationary or nonstationary. Neuroscience experiments typically involve multivariate time series data from several epochs, with the assumption that in each epoch there exists a certain number of latent stationary sources. Realistically, the dimension of these latent stationary sources should be allowed to change across epochs thereby making the overall analysis challenging. Motivated by such experiments, we develop a method to compare the spread of spectral information in several multivariate stationary processes with different dimensions. A statistic, blind to the dimension of the stationary process, is proposed to capture the proportion of spectral information in various frequency ranges and its asymptotic properties are derived. We discuss an application of the proposed method in discriminating local field potential of rats recorded before and after the occurrence of an induced stroke.

C0273: **Non-stationary high dimensional time series networks for brain imaging data**
Presenter: **Ivor Cribben**, Alberta School of Business, Canada

Original statistical methodology on the evolving interdependencies between high-dimensional multivariate time series is developed. Specifically, we introduce a data-driven method which detects change points in the network summary statistics of a (very high dimensional) multivariate time series, with each component of the time series represented by a node in the network. The novel method allows for estimation of both the time of change in the network summary statistics without prior knowledge of the number or location of the change points. We also propose a new multiple change point algorithm that begins by segmenting the data into partitions and then looks for changes locally. We show the improvement of our method over classical binary segmentation methods. We apply these methods to various simulated high dimensional data sets as well as to a resting state functional magnetic resonance imaging (fMRI) data set from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. The method allows us to characterize the large scale resting state dynamic brain networks that are related to Alzheimer's disease.

C0289: **Fast and efficient construction of estimating equation in high-dimensions by L1-truncation**

Presenter: **Davide Ferrari**, University of Bolzano, Italy

The growth in size and complexity of modern data challenges the applicability of traditional likelihood-based inference, with difficulties related to model selection and computational intractability of the full likelihood. Composite likelihood (CL) estimation avoids this by combining a number of low-dimensional likelihood objects into a single objective function used for inference. We develop a new procedure for simultaneous selection of sub-likelihood objects from a large set of feasible candidates and parameter estimation. We propose to obtain sparse CL estimating equations by minimizing the estimated distance from the full likelihood subject to a constraint representing the afforded computing cost. The resulting CL is sparse since it contains a relative small number of informative sub-likelihoods while noisy or redundant components are neglected. The new procedure is implemented by a fast least-angle algorithm and is illustrated through numerical examples.

CG061 **Room Business center** **HIGH DIMENSIONAL DATA** **Chair: Stephane Girard**

C0277: **Fast and highly efficient pseudo-likelihood methodology for large and complex ordinal data**

Presenter: **Anna Ivanova**, KU Leuven and UHasselt, Belgium

Co-authors: Geert Molenberghs, Geert Verbeke

When it comes to modelling, ordinal responses have received less attention in the literature as it considered to be complex and the latter increases further in the case joint multivariate modelling. An additional problem is the size of the collected data. Pseudo-likelihood based methods for pairwise fitting, for partitioned samples and, as introduced in our work, a combination thereof, enabled us to jointly model large amounts of responses. To illustrate our methodology, we used the composite endpoint on an ordinal scale taken from a diabetes study: multiple targets for HbA1C, LDL-cholesterol and SBP. For every response, a univariate proportional odds mixed model was formulated. To capture the association between the responses, an assumption about the correlation between random effects was made. Pseudo-likelihood methods yield valid estimates with high efficiency, even for low numbers of quadrature points. The big advantage of these methods is their gain in computation time over the full likelihood method: from 7 minutes 13 seconds (for full likelihood method) to only 20 seconds (for the combined method). This was because the submodels could run in parallel.

C0266: **Using functions of inverse means and medians for dimension reduction**

Presenter: **Andreas Artemiou**, Cardiff University, United Kingdom

Two methods for dimension reduction in regression in the sufficient dimension reduction framework are introduced. The first method is using functions of sliced inverse means for dimension reduction using two different algorithmic approaches. One algorithm is computationally much faster than existing methodology. The other algorithm performs robustly in the case of discrete response variables (i.e. iris data), that is the ordering of the data does not affect the results. To robustify against outliers we present also a method that uses functions of inverse medians.

C0263: **Variational Bayesian inference for high dimensional factor copulas**

Presenter: **Hoang Nguyen**, Universidad Carlos III de Madrid, Spain

Co-authors: Concepcion Ausin, Pedro Galeano

Factor copula models have been recently proposed for describing the joint distribution of a large number of variables in terms of a few common latent factors. However, the estimation of factor copula models in high dimensions is a challenging problem. Another issue of factor copula models is that the bivariate copula functions connecting the variables are usually unknown. We implement a Bayesian procedure to make fast inferences for multi-factor and structured factor copulas. To deal with the high dimensional structure, we apply a variational inference (VI) algorithm to estimate different specifications of factor copula models. Compared to the Markov chain Monte Carlo (MCMC) approach, the variational approximation is much faster and could handle a sizeable problem in a few seconds. We also derive an automatic procedure to recover the hidden dependence structure. By taking advantage of the posterior modes of the latent variables, we select the bivariate copula functions based on minimizing the Bayesian information criterion (BIC). We illustrate the proposed methodology with two high dimensional real data sets. The first one considers the daily temperature time series at 479 stations in Germany, while the second one analyzes the stock return dependence of 218 European companies. In general, the structured factor copula model can capture quite well the dependence structure of high dimensional data.

Monday 15.04.2019 11:30 - 13:00 Parallel Session I – CRONOSMDA2019

CI011 Room Triton 1+2 CRoNoS Session II Chair: Florian Frommlet

C0177: **Optimal tuning of subsampling Hamiltonian Monte Carlo**
Presenter: **Mattias Villani**, Linkoping University, Sweden
Co-authors: Robert Kohn, Minh-Ngoc Tran, Matias Quiroz, Doan Khue Dung Dang

Hamiltonian Monte Carlo (HMC) is an increasingly popular simulation algorithm for Bayesian inference which has proven to be especially suitable in high-dimensional problems. A drawback of HMC is that it requires a large number of evaluations of the posterior gradient, which can be computationally costly, particularly in problems with large datasets. Results on accelerating HMC by data subsampling and how to optimally tune the algorithm are presented.

C0225: **Harmonic networks and learning activation functions**
Presenter: **Rozenn Dahyot**, Trinity College Dublin, Ireland

Deep Neural Networks (DNNs) have proven very successful in many signal processing applications in the past decade. DNNs benefit from well supported and hardware optimized library for training their parameters from large amount of data and for testing. A current research trend is in revisiting older formulations or algorithms (e.g. Optimal transport, partial differential equations, wavelet analysis) using DNN formulation leading to several advantages: one is to take advantage of well optimized library; another is to provide an understanding and control of DNNs for their improvement. We present several algorithms expressed with DNNs as layers of neuron function (linear projection followed by a nonlinear activation function) with applications to colour mapping, image restoration and image classification and show that alternative strategies are emerging with pre-set linear functions combined with learned activation functions, providing insights for DNNs design and optimization.

C0234: **Computational strategies for regression subset selection**
Presenter: **Cristian Gatu**, Alexandru Ioan Cuza University of Iasi, Romania
Co-authors: Marc Hofmann, Ana Colubi, Erricos John Kontoghiorghes

Computationally efficient algorithms to compute the regression subsets are presented. They are based on regression trees and employ branch-and-bound techniques and heuristics strategies. The main numerical tool that has been employed is the QR factorization and its modification. This yields in a numerically stable and efficient sub-model estimation procedure. An R package "lmSubsets" for regression subset selection is presented. The package aims to provide a versatile tool for subset regression. It also embeds a novel algorithm that selects the best variable-subset model according to a pre-determined search criterion. This performs considerably faster than all-subsets variable selection algorithms that rely on the residual sum of squares only. Further computational time improvements based on a parallel version of the branch-and-bound are discussed.

C0238: **High performance data analytics and some applications**
Presenter: **Nahid Emad**, University of Versailles, France

In most areas of science, data production is now faster than compute capabilities. The computational modeling and data analysis associated with high performance computing techniques are used to make effectively "talk" these huge amounts of data. We highlight some challenges in the ecosystem defined by interactions between modeling, simulation and high performance data analysis. We will then see how these challenges are addressed through several examples such as Gamma ray detection in astronomy, clustering in a network of individuals or the detection of behavior anomaly in information exchange systems.

CI088 Room Triton 3 Spring course session VII Chair: Alastair Young

C0155: **Selective inference**
Presenter: **Alastair Young**, Imperial College London, United Kingdom

Selective inference is concerned with performing valid statistical inference when the questions being addressed are suggested by examination of data, rather than being specified before data collection. We describe key ideas in selective inference, from both frequentist and Bayesian perspectives. In frequentist analysis, the fundamental notion is that valid inference, in the sense of control of error rates, is only obtained by conditioning on the selection event, that is, by considering hypothetical repetitions which lead to the same inferential questions being asked. The Bayesian standpoint is less clear, but it may be argued that such conditioning on the selection is required if this takes place on the parameter space as well as on the sample space. We provide an overview of conceptual and computational challenges, as well as asymptotic properties of selective inference in both frameworks, under the assumption that selection is made in a well-defined way.

CO082 Room Neptune Functional statistics for Hilbert data Chair: Alessandra Menafoglio

C0219: **The control of the false discovery rate for functional data defined over manifold domains**
Presenter: **Alessia Pini**, Universita Cattolica del Sacro Cuore, Italy
Co-authors: Niels Lundtorp Olsen, Simone Vantini

Inference for functional data can be approached by either a global or a local perspective. In the case of local inference, a single test is performed on the entire domain. In the case of local inference instead, a p-value function is defined, assigning a p-value to each point of the domain, in order to select the portions of it responsible for the rejection of a null hypothesis. In the local setting, it is straightforward to compute a p-value at every point of the domain, obtaining an unadjusted p-value function, which controls only pointwise the probability of type I error. However, one of the main issues in this framework is how to efficiently adjust such function, in order to provide an error control over the entire domain. The focus is on the control of the false discovery rate (FDR). First, the classical notion of FDR is extended to functional data. Further, a continuous version of the Benjamini-Hochberg procedure is introduced, along with a definition of adjusted p-value function. Some general conditions are stated, under which the functional Benjamini-Hochber procedure provides control of FDR. The procedure is very general, and can be applied also to functional data defined in complex domains such as manifolds. Finally, the proposed method is applied to satellite measurements of Earth temperature. In detail, we aim at identifying the regions of the planet where temperature has significantly increased in the last decades.

C0221: **Super-consistent estimation of points of impact in nonparametric regression with functional predictors**
Presenter: **Dominik Liebl**, University Bonn, Germany
Co-authors: Dominik Poss, Alois Kneip

Predicting scalar outcomes using function-valued predictor variables is a classical problem in functional data analysis. In many applications, however, only specific locations or time-points of the functional predictors have an impact on the outcome. The selection of such points of impact constitutes a particular variable selection problem, since the high correlation in the functional predictors violates the basic assumptions of existing high-dimensional variable selection procedures. We introduce a nonparametric regression model with functional predictors evaluated at unknown points of impact which need to be estimated from the data. We propose a threshold-based and a fully data-driven estimator and derive the convergence rates of our point of impact estimators. The finite sample properties of our estimators are assessed by means of a simulation study. Our methodology is motivated by a psychological case study in which the participants were asked to continuously rate their emotional state while watching an affective online video on the persecution of African albinos.

C0233: **Weighting in Bayes spaces and its effects in statistical processing of density functions**
Presenter: **Renata Talska**, Palacky University Olomouc, Czech Republic
Co-authors: Alessandra Menafoglio, Karel Hron, Juan Jose Egozcue, Javier Palarea-Albaladejo

Probability density functions (PDFs) can be viewed as functional data carrying relative information. The relative nature of PDFs is accounted for in Bayes spaces with Hilbert space structure, which result from generalization of the Aitchison geometry for compositional data to the infinite dimensional setting. Specifically, if the focus is on PDFs restricted to a bounded support $I \subset R$, which is typically used in practical applications, they can be represented with respect to the Lebesgue reference measure within the Bayes space of positive real functions with square-integrable logarithm. The reference measure can be changed and it induces a weighting effect on the domain I. Moreover, the weighting also impacts the geometry of the Bayes spaces and results in so-called weighted Bayes spaces. The aim of this contribution is show the effects of changing the reference measure from the Lebesgue measure to a general probability measure focusing on its practical implications for the Simplicial Functional Principal Component Analysis (SFPCA). Furthermore, a centered log-ratio transformation as an isometric map between the weighted Bayes space and an unweighted L^2 space (i.e. with Lebesgue reference measure) is proposed, which enables statistical processing of PDFs using standard statistical tools provided by Functional Data Analysis (FDA).

C0236: **Prediction of spatial functional random processes: Comparing functional and spatio-temporal kriging approaches**
Presenter: **Sara Sjostedt-de Luna**, Umea University, Sweden
Co-authors: Johan Strandberg, Jorge Mateu

Functional and spatio-temporal (Sp.T.) kriging approaches to predict spatial functional random processes (which can also be viewed as Sp.T. random processes) are compared. Comparisons with respect to computational time and prediction performance via functional cross-validation are evaluated, mainly through a simulation study, but also on a real data set. We restrict comparisons to Sp.T. kriging versus ordinary kriging for functional data (OKFD), since the more flexible functional kriging approaches pointwise functional kriging (PWFK) and the functional kriging total model coincide with OKFD in several situations. We formulate conditions under which we show that OKFD and PWFK coincide. From the simulation study, it is concluded that the prediction performance of the two kriging approaches in general is rather equal for stationary Sp.T. processes. However, functional kriging tends to perform better for small sample sizes, while Sp.T. works better for large sizes. For non-stationary Sp.T. processes, with a common deterministic time trend and/or time varying variances and dependence structure, OKFD performs better than Sp.T. kriging irrespective of the sample size. For all simulated cases, the computational time for OKFD was considerably lower compared to those for the Sp.T. kriging methods.

CO054 Room Business center DATA DEPTH FOR NON-STANDARD DATA AND ITS APPLICATIONS Chair: Pavlo Mozharovskyi

C0159: **Depth for curve data and applications**
Presenter: **Myriam Vimond**, ENSAI, France

Statistical data depth has been defined as a function that determines centrality of an arbitrary point with respect to a data cloud or to a probability measure. During the last decades, this seminal idea of data depth evolved to a powerful machinery proving to be useful in various fields of science. Recently, extending the notion of data depth to the functional setting attracted a lot of attention among theoretical and applied statisticians. We go further and suggest a notion of data depth suitable for data represented as curves, or trajectories, which is independent of the parametrization. We show that our curve depth satisfies theoretical requirements of general depth functions that are meaningful for trajectories. We apply our methodology to diffusion tensor brain images and also to pattern recognition of hand written digits and letters.

C0220: **Parameter depths induced by statistical functionals and their application to control charts**
Presenter: **Ignacio Cascos**, Universidad Carlos III de Madrid, Spain

A parameter depth function assesses how well does an element of a parameter space fit a probability distribution as its parameter (of some given kind). Specifically, we define a parameter depth of an element of a parameter space (candidate) induced by a statistical functional with respect to a probability distribution as one minus the smallest fraction of mass that needs to be blurred away from the distribution so that the candidate matches the statistical functional evaluated on the remaining mass. If such statistical functional is the mean, then we end up with a notion of data depth (centrality of points with respect to a distribution), precisely the zonoid depth. Several parameter depth notions are proposed, and used as charting statistics in control charts. These charts evaluate how well does the estimate of a parameter on a rational sample fit some historical dataset in terms of the parameter depth with respect to it. They have thus a unique control limit which is established either by Monte Carlo or bootstrap techniques.

C0226: **Distance based depth functions for directional data**
Presenter: **Giuseppe Pandolfo**, University of Naples Federico II, Italy
Co-authors: Davy Paindaveine, Giovanni Camillo Porzio

Data depth are aimed at providing an inner-outer ordering of data in multivariate spaces with respect to a sample or a distribution. Depth functions have also been defined on spheres and some notions are available within the literature. However, these either lack flexibility or are very computationally expensive and thus can only be used in small dimensions. Hence, we introduce a class of depth functions for directional data which are based on spherical distances. These depths are computationally feasible also in high dimensions. Structural properties of the proposed depths are derived along with the asymptotic and robustness properties of the corresponding deepest points. The practical relevance of the proposed depths are shown through two potential applications in directional statistics, related to (i) spherical location estimation and (ii) supervised classification.

C0160: **Nonparametric imputation by data depth**
Presenter: **Pavlo Mozharovskyi**, Telecom ParisTech, Paris Saclay University, France
Co-authors: Julie Josse, Francois Husson

A single imputation method for missing values is presented which borrows the idea of data depth - a measure of centrality defined for an arbitrary point of a space with respect to a probability distribution or data cloud. This consists in iterative maximization of the depth of each observation with missing values, and can be employed with any properly defined statistical depth function. For each single iteration, imputation reverts to optimization of quadratic, linear, or quasiconcave functions that are solved analytically, by linear programming or the Nelder-Mead method. As it accounts for the underlying data topology, the procedure is distribution free, allows imputation close to the data geometry, can make prediction in situations where local imputation (k-nearest neighbors, random forest) cannot, and has attractive robustness and asymptotic properties under elliptical symmetry. It is shown that a special case - when using the Mahalanobis depth - has direct connection to well-known methods for the multivariate normal model, such as iterated regression and regularized PCA. The methodology is extended to multiple imputation for data stemming from an elliptically symmetric distribution. Simulation and real data studies show good results compared with existing popular alternatives. The method has been implemented as an R-package.

Monday 15.04.2019 14:30 - 16:30 Parallel Session J – CRONOSMDA2019

CI090 Room Triton 3 SPRING COURSE SESSION VIII Chair: Alastair Young

C0276: **Selective inference**

Presenter: **Alastair Young**, Imperial College London, United Kingdom

Selective inference is concerned with performing valid statistical inference when the questions being addressed are suggested by examination of data, rather than being specified before data collection. We describe key ideas in selective inference, from both frequentist and Bayesian perspectives. In frequentist analysis, the fundamental notion is that valid inference, in the sense of control of error rates, is only obtained by conditioning on the selection event, that is, by considering hypothetical repetitions which lead to the same inferential questions being asked. The Bayesian standpoint is less clear, but it may be argued that such conditioning on the selection is required if this takes place on the parameter space as well as on the sample space. We provide an overview of conceptual and computational challenges, as well as asymptotic properties of selective inference in both frameworks, under the assumption that selection is made in a well-defined way.

CO044 Room Neptune INFERENCES ABOUT MULTIVARIATE AND FUNCTIONAL DATA Chair: Sophie Dabo

C0179: **Estimation of a partially linear additive model with generated covariates**

Presenter: **Carlos Martins-Filho**, University of Colorado at Boulder, United States

Co-authors: Xin Geng, Feng Yao

Kernel-based estimators are proposed for both the parametric and nonparametric components of a partially linear additive regression model where a subset of the covariates entering the nonparametric component are generated by the estimation of an auxiliary nonparametric regression. Both estimators are shown to be asymptotically normally distributed. The estimator for the finite dimensional parameter is shown to converge at the parametric square-root-n rate and the estimator for the infinite dimensional parameter converges at a slower nonparametric rate that, as usual, depends on the rate of decay of the bandwidths and the dimensionality of the underlying regression. A small Monte Carlo study is conducted to shed light on the finite sample performance of our estimators and to contrast them with those of estimators available in the extant literature.

C0181: **Nonparametric inference for copulas and measures of dependence under length-biased sampling and informative censoring**

Presenter: **Taoufik Bouezmarni**, Universite de Sherbrooke, Canada

Length-biased data are often encountered in cross-sectional surveys and prevalent-cohort studies on disease durations. Under length-biased sampling subjects with longer disease durations have greater chance to be observed. As a result, covariate values linked to the longer survivors are favoured by the sampling mechanism. When the sampled durations are also subject to right censoring, the censoring is informative. Modelling dependence structure without adjusting for these issues leads to biased results. We consider copulas for modelling dependence when the collected data are length- biased and account for both informative censoring and covariate bias that are naturally linked to length-biased sampling. We address nonparametric estimation of the bivariate distribution, copula function and its density, and Kendall and Spearman measures for right-censored length-biased data. The proposed estimator for the bivariate cdf is a Hadamard- differentiable functional of two MLEs (Kaplan-Meier and empirical cdf) and inherits their efficiency. Based on this estimator, we devise two estimators for copula function and a local-polynomial estimator for copula density that accounts for boundary bias.

C0203: **Partially linear spatial probit models**

Presenter: **Mohamed Salem Ahmed**, University of Lille, France

Co-authors: Sophie Dabo, Michael Genin

Partially linear probit model for spatially dependent data is considered. A triangular array setting is used to cover various patterns of spatial data. Conditional heteroscedasticity and non-identically distributed observations and a linear process for disturbances are assumed allowing various spatial dependencies. The estimation procedure proposed is a combination of a weighted likelihood and a generalized method of moments. We first fix the parametric components of the model and estimate the nonparametric one using weighted likelihood. The obtained estimate is then used to construct a GMM parametric component estimate. Consistency of the parametric and non-parametric components estimators and asymptotic normality of the parameter one are established under sufficient conditions. We present some simulated experiments including real data to investigate the finite sample performance of the estimators.

C0250: **Estimating a covariance function from fragments of functional data**

Presenter: **Aurore Delaigle**, University of Melbourne, Australia

Functional data are often observed only partially, in the form of fragments. In that case, the standard approaches for estimating the covariance function do not work because entire parts of the domain are completely unobserved. Previously we have suggested ways of estimating the covariance function, based for example on Markov assumptions. We take a completely different approach which does not rely on such assumptions. We show that, using a tensor product approach, it is possible to reconstruct the covariance function using observations located only on the diagonal of its domain.

C0254: **Quasi-maximum likelihood estimators for functional linear spatial autoregressive models**

Presenter: **Zied Gharbi**, University of Lille, France

Co-authors: Sophie Dabo

Spatial functional random variables are becoming more common in statistical analyses due to the availability of high-frequency spatial data and new mathematical strategies to address such statistical objects particularly within the scope of geostatistics. In the geostatistical approach, spatial locations are continuous compare to many domains such as remote sensing from satellites, image analysis, weather patterns, agriculture and so on, where spatial domains are counties or census tracts or in general received as regular lattice. The main focus is the lattice setting where we propose a functional linear autoregressive spatial model, where the explanatory variable takes values in a functional space while the response process is real-valued and spatially autocorrelated. The particularity of the model is due to the functional nature of the explanatory variable and the use of a spatial weight matrix that defines the spatial dependency between neighbors. The estimation procedure consists of reducing the infinite dimension of the functional explanatory variable and maximizing a quasi-maximum likelihood. We establish the consistency and asymptotic normality of the estimator. Numerical experiments by simulations and an application to real data are given.

CO023 **Room Business center** **ADVANCES IN FINANCIAL ECONOMETRICS** **Chair: Erricos Kontoghiorghes**

C0154: **Using text mining methods for comparing trends in topics in economic journals overtime**
Presenter: **Peter Winker**, University of Giessen, Germany
Co-authors: David Lenz

The comparison of information content of different sources or within the same source over time is highly relevant. We compare text corpora consisting of articles published in two economic journals. Thereby, the main focus is on the development of topic importance over time and how it (co-)evolves. We present a quantitative framework for comparing text corpora based on their latent topics using text mining techniques. Time information is utilized to track the evolution of the relevance of these topics. We evaluate three comparison methods: Treat both text corpora as a single corpus, train a model on one corpus and evaluate the other corpus based on this model and vice versa, and train a model for each corpus and use a matching approach for pairing corresponding topics. For the empirical application, we exploit the corpus of articles published in the Journal of Economics and Statistics and the corpus of articles published in the Review of World Economics, both from 1913 to 1941. We present topic dynamics for both corpora and their (co-)evolution over time. Furthermore, the analysis indicates which of the methods is most promising for this type of analysis.

C0169: **On the frequency of transmission of market volatility: A double asymmetric GARCHMIDAS approach**
Presenter: **Alessandra Amendola**, Department of Economics and Statistics - University of Salerno, Italy
Co-authors: Vincenzo Candila, Giampiero Gallo

Volatility in financial markets has both low and high-frequency components which determine its dynamic evolution. Previous modelling efforts in the GARCH context (e.g. the SplineGARCH) were aimed at estimating the low frequency component as a smooth function of time around which short-term dynamics evolves. Alternatively, recent literature has introduced the possibility of considering data sampled at different frequencies to estimate the influence of macrovariables on volatility. We propose to use an extension of the GARCH MIDAS model, called Double Asymmetric GARCH MIDAS, where variations in a market volatility variable are observed both at the daily and the monthly level and represent different channels through which market volatility can influence individual stocks. We want to convey the idea that such variations (separately) affect the short and longrun components, possibly having a separate impact according to their sign.

C0272: **Independent and conditionally independent counterfactuals**
Presenter: **Marcin Wolski**, European Investment Bank, Luxembourg

A novel dependence filtering framework is proposed. A counterfactual random variable which is independent, or conditionally independent, from the effects of given covariates is characterized. Under error exogeneity the counterfactuals have causal interpretation. A fully nonparametric estimation technique and an inference roadmap for such counterfactuals is offered. A numerical exercise confirms that the approach performs well in nonlinear environments. Furthermore, bootstrap validity results are provided for the confidence intervals of the estimates. The approach is applied to filter out the sovereign risk spill-overs on corporate cost of borrowing in selected euro area countries.

C0150: **High frequency linear time series models and mixed frequency data**
Presenter: **Manfred Deistler**, Vienna University of Technology, Austria

The focus is on the identification of multivariate linear dynamic models from so called mixed frequency (MF) data, i.e. from data where the univariate components of the time series are sampled at different frequencies; in economic applications this occurs if e.g. unemployment data are sampled monthly and GNP is available only quarterly. The interest is in the underlying “high frequency” (HF) model, i.e. in the model generating outputs at the highest sampling frequency. The model classes considered are multivariate AR and ARMA models (both with nonsingular and singular innovation variance) and linear dynamic factor models. We discuss problems of parameter identifiability and of estimation. In estimation in particular MLEs and EM algorithms are analyzed, both w.r.t their asymptotic and finite sample properties. The information loss due to MF data relative to HF data is discussed.

C0297: **IDEOLOG: A program for filtering econometric data**
Presenter: **Stephen Pollock**, University of Leicester, United Kingdom

The IDEOLOG program originated in the desire to compare some new methods of filtering with existing procedures that are common in econometric analysis. The outcome has been a comprehensive facility that will enable a detailed analysis of univariate econometric time series, as well as time series originating from many other investigations. The program will serve to reveal the extent to which the results of an economic analysis might be the consequence of the choice of a particular filter.

CO078 **Room Triton 1+2** **RECENT DEVELOPMENTS IN CLUSTERING AND CLASSIFICATION METHODS** **Chair: Marta Nai Ruscone**

C0223: **Clustering finite mixture components**
Presenter: **Roberto Rocci**, University of Rome Tor Vergata, Italy

Finite mixture of Gaussians is a well-known model frequently used to classify a sample of observations. The idea is to consider the sample as drawn from a heterogeneous population where each sub-population, or group, is described by a component of the mixture. This is correct if the hypothesis that each group has a multivariate normal distribution is true. Otherwise, the estimated number of components G can be greater than the true number of subpopulations because the distribution of a subpopulation can be well approximated by a finite mixture of Gaussians. In this case, two or more components correspond to the same subpopulation, but we do not know who they are. We try to overcome the aforementioned problem by assuming that the population is formed by K subpopulations, each having a distribution that can be described, or well approximated, by a finite mixture of Gaussians. The consistent estimation of this model is not possible if the labels of the K groups are unknown, because it is not identified. Therefore, we approach the problem by first fitting a mixture of G Gaussians to the data and then by combining the components into K ($< G$) groups by maximizing an appropriate criterion.

C0187: **Variable selection for model-based clustering**
Presenter: **Matthieu Marbac**, CREST - ENSAI, France
Co-authors: Mohammed Sedki

Two approaches are presented for selecting variables in latent class analysis. The first approach consists in optimizing the BIC with a modified version of the EM algorithm. This approach simultaneously performs both model selection and parameter inference. The second approach consists in maximizing the MICL, which considers the clustering task, with an algorithm of alternate optimization. This approach performs model selection without requiring the maximum likelihood estimates for model comparison, then parameter inference is done for the unique selected model. Thus, both approaches avoid the computation of the maximum likelihood estimates for each model comparison. Moreover, they also avoid the use of the standard algorithms for variable selection which are often suboptimal (e.g. stepwise method) and computationally expensive. The case of data with missing values is also discussed. The interest of both proposed criteria is shown on an application in human population genomics problem. Data set describes 1300 patients by 160000 variables.

C0173: **Dealing with missing data in model-based clustering through a MNAR model**

Presenter: **Christophe Biernacki**, Inria, France

Co-authors: Gilles Celeux, Julie Josse, Fabien Laporte

Since the 90s, model-based clustering is largely used to classify data. Nowadays, with the increase of available data, missing values are more frequent. Traditional ways to deal with them consist in obtaining a filled data set, either by discarding missing values or by imputing them. In the first case some information is lost; in the second case the final clustering purpose is not taken into account through the imputation step. Thus, both solutions risk blurring the clustering estimation result. Alternatively, we defend the need to embed the missingness mechanism directly within the clustering modeling step. There exist three types of missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). In all situations, logistic regression is proposed as a natural and flexible candidate model. In particular, its flexibility property allows us to design some meaningful parsimonious variants, as dependency on missing values or dependency on the cluster label. In this unified context, standard model selection criteria can be used to select between such different missing data mechanisms, simultaneously with the number of clusters. Practical interest of our proposal is illustrated on data derived from medical studies suffering from many missing data.

C0171: **Clustering ranking data via copulas**

Presenter: **Marta Nai Ruscone**, LIUC, Italy

Clustering of ranking data aims at the identification of groups of subjects with a homogenous, common, preference behavior. Ranking data occurs when a number of subjects are asked to rank a list of objects according to their personal preference order. The input in cluster analysis is a distance matrix, whose elements measure the distances between rankings of two subjects. The choice of the distance dramatically affects the final result and therefore the computation of an appropriate distance matrix is an issue. Several distance measures have been proposed for ranking data. The most important are the Kendalls t, Spearmans r and Cayley distances. When the aim is to emphasize top ranks, weighted distances for ranking data should be used. We propose a generalization of this kind of distances using copulas. Those generalizations provide a more flexible instrument to model different types of data dependence structures and consider different situations in the classification process. Simulated and real data are used to illustrate the pertinence and the importance of our proposal.

C0201: **Spatial clustering of average risks and risk trends in Bayesian disease mapping**

Presenter: **Nema Dean**, University of Glasgow, United Kingdom

Disease mapping is the study of variability in disease risk typically across space but also over time. It is to be expected that the variation in these risks will be due to environmental, geographical and risk inducing behaviours - spatial effects - as well temporal changes in these factors, for example due to new legislation being enforced. A Bayesian approach is discussed to combining both spatial and trend clustering in the context of disease mapping for areal data where aggregated information is available on a partition of a region rather than individual level information. The goal is to group areas according to both similar average level of risk and also risk trends. However, it is assumed that areas close together are likely to have similar behaviour from a disease risk point of view and a conditional autoregressive prior is used to enforce smoothness within but not between clusters. The effectiveness of this model is examined through simulations and illustrated on a study of respiratory disease risk in Glasgow, Scotland.

Monday 15.04.2019 17:00 - 18:30 Parallel Session K – CRONOSMDA2019

CI009 Room Triton 1+2 CRoNoS SESSION III Chair: Alastair Young

C0196: **Quantifying and estimating asymmetric dependence**
Presenter: **Wolfgang Trutschnig**, University of Salzburg, Austria

Standard dependence measures considered in the (mostly non-mathematical) literature like Pearson correlation, Spearman correlation, and Schweitzer and Wolff's famous Σ are symmetric, i.e. they assign each pair (X,Y) of random variables the same dependence as they assign the pair (Y,X). Independence of two random variables is a symmetric concept modelling the situation that knowing X does not change our knowledge about Y and vice versa - dependence, however, is not. Thinking, for instance, of a sample $(x_1,y_1),\ldots,(x_n,y_n)$ roughly in the shape of a noisy letter V, it is without doubt (on average) easier to predict the y-value given the x-value than vice versa. The R-package qad (short for quantification of asymmetric dependence) aims at detecting asymmetries in samples. It estimates the dependence of the second variable on the first one and vice versa, and additionally quantifies the asymmetry of the underlying dependence structure. The main objectives are to sketch the basic ideas behind qad, to present the most relevant mathematical properties of the underlying estimator(s), and to illustrate its capabilities by some examples.

C0217: **Donsker results for the smoothed empirical process of dependent observations**
Presenter: **Eric Beutner**, Maastricht University, Netherlands
Co-authors: Henryk Zaehle

It is well known that standard kernel based density estimators are mean integrated squared error optimal for an appropriate bandwidth but fail to be square root n consistent when used to estimate, for instance, moments of the underlying distribution or more generally a functional of the underlying distribution. We focus on kernels and bandwidths that lead to mean integrated squared error optimal estimators and that are, at the same time, square root n consistent for large classes of functionals. It will be demonstrated that the results do not only hold in the iid setting but also for many stationary time series models.

C0182: **New foundations for functional local linear regression**
Presenter: **Frederic Ferraty**, Mathematics Institute of Toulouse, France
Co-authors: Stanislav Nagy

Local linear regression is one of the most popular nonparametric regression method when the predictor is a finite-dimensional covariate. It is well known that the local linear regression outperforms the usual kernel estimator and the literature dealing with this topic is huge. To our knowledge (and surprisingly) there are only two papers extending the local linear regression to the situation when one considers a functional predictor. Problem: the theoretical developments of one of these works is approximative where in the second one, the authors require strong assumptions with respect to the distribution of the functional predictor. Even if the infinite-dimensional feature of the predictor makes challenging the asymptotics in the functional local linear regression, it is clear that this topic is still underdeveloped. The aim is to bring a relevant response by proposing new theoretical developments. As a by-product, we also provide the asymptotics for the Frechet derivative of the functional linear regression operator.

C0153: **On the design of experiments with ordered treatments**
Presenter: **Ori Davidov**, University of Haifa, Israel

There are many situations where one expects an ordering among $K \geq 2$ experimental groups or treatments. Although there is a large body of literature dealing with the analysis under order restrictions, surprisingly, very little work has been done in the context of the design of experiments. A principled approach to the design of experiments with ordered treatments is provided. In particular we propose two classes of designs which are optimal for testing different types of hypotheses. The theoretical findings are supplemented with thorough numerical experimentation and a concrete data example. It is shown that there is a substantial gain in power, or alternatively a reduction in the required sample size, when an experiment is both designed and analyzed using methods which account for order restrictions.

CI092 Room Triton 3 SPRING COURSE SESSION IX Chair: Peter Winker

C0278: **Text mining in econometrics**
Presenter: **Peter Winker**, University of Giessen, Germany

There is a growing interest in the use of textual information in different fields of economics ranging from financial markets (analysts statements, communication of central banks) over innovation activities (patent abstracts, websites) to the history of economic science (journal articles). In order to draw meaningful conclusions from this type of data, the analysis has to cover a substantial number of steps including 1) the selection of appropriate sources (corpora) and establishing access, 2) the preparation of the text data for further analysis, 3) the identification of themes within documents, 4) quantifying the relevance of themes in different documents, 5) aggregating relevance information, e.g. across sectors or over time, 6) analysis of the generated indicators. The course will provide some first insights into these steps of the analysis and indicate open issues regarding, e.g. computational complexity and robustness of the methods. It will be illustrated with empirical examples.

CO042 Room Neptune ROBUST METHODS FOR HIGH DIMENSIONAL DATA Chair: Stefan Van Aelst

C0207: **The minimum regularized covariance determinant estimator**
Presenter: **Tim Verdonck**, KU Leuven, Belgium
Co-authors: Peter Rousseeuw, Kris Boudt, Steven Vanduffel

The Minimum Covariance Determinant (MCD) approach estimates the location and scatter matrix using the subset of given size with lowest sample covariance determinant. Its main drawback is that it cannot be applied when the dimension exceeds the subset size. We propose the Minimum Regularized Covariance Determinant (MRCD) approach, which differs from the MCD in that the scatter matrix is a convex combination of a target matrix and the sample covariance matrix of the subset. A data-driven procedure sets the weight of the target matrix, so that the regularization is only used when needed. The MRCD estimator is defined in any dimension, is well-conditioned by construction and preserves the good robustness properties of the MCD. We prove that so-called concentration steps can be performed to reduce the MRCD objective function, and we exploit this fact to construct a fast algorithm. We verify the accuracy and robustness of the MRCD estimator in a simulation study and illustrate its practical use for outlier detection and regression analysis on real-life high-dimensional data sets in chemistry and criminology.

C0189: **Fast robust correlation for high dimensional data**
Presenter: **Jakob Raymaekers**, KULeuven, Belgium
Co-authors: Peter Rousseeuw

The product moment covariance is a cornerstone of multivariate data analysis, from which one can derive correlations, principal components, Mahalanobis distances and many other results. Unfortunately the product moment covariance and the corresponding Pearson correlation are very susceptible to outliers (anomalies) in the data. Several robust measures of covariance have been developed, but few are suitable for the ultrahigh dimensional data that are becoming more prevalent nowadays. For that one needs methods whose computation scales well with the dimension, are guaranteed to yield a positive semidefinite covariance matrix, and are sufficiently robust to outliers as well as sufficiently accurate in the statistical sense of low variability. We construct such methods using data transformations. The resulting approach is simple, fast and widely applicable. We study its robustness by deriving influence functions and breakdown values, and computing the mean squared error on contaminated data. Using these results we select a method that performs well overall, which we call wrapping. It is illustrated on genomic data with 12,000 variables and color video data with 920,000 dimensions.

C0195: **Robust inference and modeling of mean and dispersion for generalized linear models**
Presenter: **Jolien Ponnet**, KU Leuven, Belgium
Co-authors: Pieter Segaert, Stefan Van Aelst, Tim Verdonck

Generalized linear models (GLMs) are a very popular class of regression models when the responses follow a distribution in the exponential family. In real data the variability often deviates from the relation imposed by the exponential family distribution, which results in over- or underdispersion. Dispersion effects may even vary in the data. Such data sets do not follow the traditional GLM distributional assumptions anymore, leading to unreliable inference. Therefore, the family of double exponential distributions has been proposed, which allows to model both the mean and the dispersion as a function of covariates in the GLM framework. However, it is well known that standard maximum likelihood inference is highly susceptible to the possible presence of outliers. To overcome this pitfall, the robust double exponential (RDE) estimator is proposed. Simulation studies for binomial and Poisson models show the good performance of the RDE estimator in comparison to existing alternatives. Applications to real data illustrate the relevance of robust inference for dispersion effects in GLMs.

C0211: **Robust estimation for functional and partially functional linear models**
Presenter: **Ioannis Kalogridis**, KU Leuven, Belgium
Co-authors: Stefan Van Aelst

Functional data analysis is a fast evolving branch of modern statistics, yet despite the popularity of the functional linear model in recent years, current estimation procedures either suffer from lack of robustness or are computationally burdensome. To address these drawbacks, we propose a flexible family of lower-rank smoothers that combines penalized splines and M-estimation. We show that, under an additional condition on the design matrix, these estimators exhibit the same asymptotic properties as the corresponding least-squares estimators, while being considerably more reliable in the presence of outliers. Further, the proposed methods easily generalize to functional models that include scalar covariates or nonparametric components, thus providing a wide framework of estimation. Simulation experiments show that the proposed estimators have high efficiency, protect against vertical outliers, produce smooth estimates and compare favourably with existing procedures, least-squares and robust alike.

CC067 Room Business center STATISTICAL MODELLING Chair: Mattias Villani

C0282: **Regression modelling with I-priors**
Presenter: **Wicher Bergsma**, London School of Economics, United Kingdom

A methodology is introduced which unifies and generalizes a variety of regression methods and models, including multilevel, varying coefficient, and longitudinal models, and models with functional covariates and/or responses. The methodology has some advantages compared to some commonly used methods in terms of ease of estimation and model comparison. Our approach is an empirical Bayes one built around the I-prior, which is an objective Gaussian process (GP) prior for a regression function in a reproducing kernel Krein space (RKKS), and is based on the Fisher information. In the regression model, each covariate, which may be multidimensional or functional, has a scale parameter which we estimate using maximum marginal likelihood. In contrast to GP priors which are subjectively chosen, the I-pior permits a simple EM algorithm for estimating these scale parameters, facilitating estimation, especially if there are many covariates. The proposed approach entails high model parsimony, with one scale parameter for each covariate, but no additional ones are needed for interaction effects. This allows a semi-Bayes approach to the selection of interaction effects, able to detect effects with smaller sample sizes than the classical approach. Finally, what we see as one of the major advantages of this unified viewpoint is that it gives the potential to make it easier for users to specify models in software. An R-package implementing our methodology is available.

C0170: **Bayesian analysis of predictive non-homogeneous hidden Markov models using Polya-Gamma data augmentation**
Presenter: **Constandina Koki**, Athens University of Economics and Business, Greece
Co-authors: Loukia Meligkotsidou, Ioannis Vrontos

Two-state Non-Homogeneous Hidden Markov Models (NHHMMs) are considered for forecasting univariate time series. The time series are modeled via different predictive regression models for each state. Also, the time-varying transition probabilities depend on exogenous variables through a logistic function. In a hidden Markov setting, inference for logistic regression coefficients becomes complicated and in some cases impossible due to convergence issues. A recently proposed latent variable scheme that utilizes the Polya-Gamma class of distributions is used to address this problem. In addition, model uncertainty regarding the predictors that affect the series both linearly -in the mean- and non-linearly -in the transition matrix- is accounted for. Predictor selection and inference on the model parameters are based on an MCMC scheme with reversible jump steps. Single-step and multiple-steps-ahead predictions are obtained based on the most probable model, the median probability model or a Bayesian Model Averaging (BMA) approach. Simulation experiments, as well as an empirical study on real financial data, illustrate the performance of our algorithm in various setups, in terms of mixing and convergence properties, model selection and predictive ability.

C0279: **Asymmetric distributions and quantile estimation**
Presenter: **Anneleen Verhasselt**, Hasselt University, Belgium
Co-authors: Irene Gijbels, Md Rezaul Karim

A general class of asymmetric distributions is studied. Their probabilistic properties lead to explicit expressions for all main characteristics (mean, variance, skewness, kurtosis, ...). Estimation of the parameters via method of moments and the maximum likelihood method is discussed, and the asymptotic behaviour of the estimators is established, again in the general framework. The emphasis in the inference is on quantile estimation. Interesting examples include new asymmetric normal, logistic and Student t distributions. The practical use of the studied asymmetric distributions is illustrated via real data examples. In a regression setting the interest is in estimating conditional quantiles. Starting from the above family of asymmetric densities, we consider a class of conditional density functions, in which the conditional quantile takes the form of a simple location-scale expression. Local likelihood techniques are then used to provide semiparametric estimates of the regression quantile curves.

C0158: **Statistical tests to identify unknown number of outliers in linear regression**
Presenter: **Linas Petkevicius**, Vilnius University, Lithuania
Co-authors: Vilijandas Bagdonavicius

Outliers identification is very important part of any stage of data analysis. More importantly, the actual number of outliers in real case situations is unknown. Usually tests are focused on a fixed number of outliers. We suggest the definition of contaminant outliers and propose a simple test to identify the unknown number of outliers. After a strict definition of contaminant outlier, new tests for multiple outliers identification in linear regression models are proposed. Approximations for the critical values of the test statistics are given. If the hypothesis of absence of outliers is rejected, then classification rules of the observations to outliers and non-outliers are given. The performance of new tests is compared with the performance of well-known tests by extensive simulations. Only the tests giving strict formal outlier classification rules are considered. In many situations the new tests give very good results from the point of view of masking and swamping values. Several well-known real data examples are analyzed using each of the considered tests and their performances are compared.

Tuesday 16.04.2019 08:50 - 10:20 Parallel Session L – CRONOSMDA2019

CI096 Room Triton 3 SPRING COURSE SESSION X Chair: Zlatko Drmac

C0287: **Numerical linear algebra for computational statistics**
Presenter: **Zlatko Drmac**, University of Zagreb, Croatia

Why has a covariance matrix negative eigenvalues and other questions related to numerical procedures in computational statistics are discussed, in particular on eigenvalues, singular value decomposition (SVD) and its generalization, the GSVD. These are the tools of trade in various applications, including computational statistics, least squares modeling, vibration analysis in structural engineering - just to name a few. In essence, the GSVD can be reduced to the SVD of certain products and quotients of matrices. For instance, in the canonical correlation analysis of two sets of variables x, y, with joint distribution and the covariance matrix $C = (C_{xx}, C_{xy}; C_{yx}, C_{yy})$, wanted is the SVD of the product $C_{xx}^{-1/2} C_{xy} C_{yy}^{-1/2}$. However, using software implementations of numerical algorithms is not that simple, despite availability of many well-known state-of-the-art software packages. We will review the recent advances in this important part of numerical linear algebra, with particular attention to *(i)* understanding the sensitivity and condition numbers; *(ii)* numerical robustness and limitations of numerical algorithms; *(iii)* careful selection and deployment of reliable mathematical software to be able to interpret and use the computed output with confidence in concrete applications. We illustrate the theoretical numerical issues on selected tasks from computational statistics.

CO060 Room Neptune ANALYSIS OF HIGH DIMENSIONAL DATA Chair: Malgorzata Bogdan

C0235: **Random covariance matrices for multivariate and high-dimensional covariance-mean mixture normal models**
Presenter: **Krzysztof Podgorski**, Lund University, Sweden

Random scaling and centering for normal random variables and vector serves as a well-known mechanism to go beyond strictly Gaussian model. However, for multivariate variables a natural extension uses a square root of a random covariance to mix random Gaussian vectors. Randomly varying dependence between coordinates in the data provides meaningful interpretation to such an approach. The matrix valued gamma distributions when used as a mixed covariances lead to gamma-covariance models. They constitute an extension of Wishard distribution, known from the theory of covariance estimation for the normal vectors. In high dimensional setting the problem of potential singularity of the underlying covariances can be addressed by a singular matrix-variate gamma distribution which is a natural extension of a singular Wishart distribution and complement the class of non-singular matrix valued gamma distribution. Their fundamental properties are discussed including densities, moments and characteristic functions. For the so extended class of matrix-variate gamma distributions we study in full detail the infinite divisibility and some related group properties both from the historical and structural perspectives and provide some new results on the topic. Finally, a matrix-variate Laplace distribution are introduced as covariance-mean mixtures of normally distributed matrices. Their potential for spatial-temporal multivariate models with locally varying covariance matrices will be outlined.

C0228: **Sparse index tracking via the sorted ℓ_1 - norm**
Presenter: **Sandra Paterlini**, University of Trento, Italy
Co-authors: Malgorzata Bogdan, Philipp Johannes Kremer, Damian Brzyski

Index tracking and hedge fund replication aims at replicating or cloning the risk-returrn properties of a given benchmark, by either using only a subset of its original constituents or by a set of risk factors. We propose a new statistical model for index tracking and hedge fund replication, that relies on the convex *Sorted ℓ_1 Penalized Estimator* (SLOPE). SLOPE is capable not only to provide sparse clones, that is replicating portfolios with few active positions, but also to automatically group assets sharing similar statistical properties with respect to the benchmark, and thereby allowing to develop further investment strategies. Considering equity index data over the period from December 2004 to January 2016 and hedge fund returns from June 1994 to July 2017, we show that our newly created tracking strategies can outperform state-of-the-art approaches in terms of tracking error, sparsity and turnover. The new method can then be considered an effective tool for quantitative asset management and robo-advisory companies.

C0264: **Effective probability distributions for spatially dependent processes**
Presenter: **Anastassia Baxevani**, University of Cyprus, Cyprus
Co-authors: Dionissios Hristopulos

Spatially distributed physical processes can be modelled as random fields with their complex spatial dependence being incorporated in their joint probability density function. Unfortunately, although most of the physical processes observed often exhibit significant deviations from the Gaussian assumption, there is only a handful of non-Gaussian joint density functions that admit explicit expressions. In addition, spatial random field models based on Gaussian or non-Gaussian joint densities incur formidable computational costs for big datasets. We propose an effective distribution approach which replaces the joint probability density with a product of univariate conditional density functions modified by a local interaction term. The effective densities involve local parameters that link the densities at different locations by means of kernel regression. We propose a sequential simulation approach by generating multiple field realisations based on effective distribution models and local constraints. The proposed method can capture well non-Gaussian dependence and are applicable to large spatial datasets, since they do not require the storage and inversion of large covariance matrices. We compare the effective distribution approach with a classical method of conditional simulation using Gaussian and non-Gaussian synthetic data.

C0227: **Empirically driven orthonormal bases for functional data analysis**
Presenter: **Hiba Nassar**, Lund University, Sweden
Co-authors: Krzysztof Podgorski

There is a strong relation between high-dimensional data and functional data. One can convert the densely observed high-dimensional data to functional data by defining a set of functional basis, then set up the coefficients to define the functional data as a linear combination of these bases. Typically, the choice of the bases is not data-driven with a notable exception to the number of dimensions, that often can be derived by cross-validations. As a consequence several standard bases such as Fourier and related bases, wavelets, splines etc. are typically used to transform observed functional data. Through such a prior and rather arbitrary decision on the basis selection the problem is transformed to a finite-dimensional space of basis coefficients and formally is losing its infinite dimensional character. We propose a strictly data-driven method of basis selection. Since the method is algorithmic and searches the data to find an effective representation of the basis by minimizing overall mean square error across functional samples, the functional basis is strictly tied to the functional character of the data and loses arbitrariness of common approaches. The method itself uses B-splines and O-splines in the machine learning style of functional data mining to find efficiently placed knots. Due to machine learning character of data processing, the method has the potential to further numerically improve and extend beyond the considered scope.

CO033 Room Business center ADVANCES IN COMPUTATIONAL STATISTICS Chair: Stella Hadjiantoni

C0198: **Estimating the number of signals in noisy ICA**

Presenter: **Klaus Nordhausen**, Vienna University of Technology, Austria
Co-authors: Joni Virta

In a noisy ICA framework the number of independent signal components is usually considered unknown. Using PCA, inferential tools for determining the number of independent component are introduced. The tools are either based on asymptotic arguments or use resampling strategies.

C0237: **Pivotal methods for clustering**

Presenter: **Roberta Pappada**, University of Trieste, Italy
Co-authors: Leonardo Egidi, Francesco Pauli, Nicola Torelli

Given a partition of statistical units into k non-overlapping groups, pivotal methods aim at identifying those units -called pivots- that are not connected with a large number of units belonging to the other groups. The pivots are selected using a similarity matrix between units, in order to choose the observations that are 'as far as possible from each other'. In terms of a co-association matrix combining multiple clusterings, the submatrix corresponding to the pivotal units will be identical or nearly identical. Pivotal methods can be adopted to deal with the label switching problem in Bayesian estimation of finite mixture models. When the number of clusters is kept small, the so-called Maxima Units Search (MUS) algorithm can be used for extracting pivotal units from a large and sparse similarity matrix. An interesting application arises in the framework of k-means clustering, whose main limitation is represented by the impact of the initial random seeding on the final solution. We propose a modification of the classical k-means algorithm which uses the information coming from clustering ensembles and the pivots in the initialization step, in order to improve the final configuration. A simulation study is presented, involving different scenarios in which the classical approach may fail to identify the 'natural' groups. Finally, the R package 'pivmet' for an approach to relabelling and k-means clustering based on MUS and/or other pivotal methods is illustrated.

C0252: **A simple recipe for making accurate parametric inference in finite sample**

Presenter: **Samuel Orso**, University of Geneva, Switzerland
Co-authors: Stephane Guerrier, Mucyo Karemera, Maria-Pia Victoria-Feser

Constructing tests or confidence regions that control over the error rates in the long-run is probably one of the most important problem in statistics. Yet, the theoretical justification for most methods in statistics is asymptotic. The bootstrap for example, despite its simplicity and its widespread usage, is an asymptotic method. There are in general no claim about the exactness of inferential procedures in finite sample. We propose an alternative to the parametric bootstrap. We setup general conditions to demonstrate theoretically that accurate inference can be claimed in finite sample.

C0269: **Sensitivity to outliers of functional depth**

Presenter: **Alicia Nieto-Reyes**, Universidad de Cantabria, Spain

Functional depth provides an order to a functional space with respect to a probability distribution or functional data set. We compare through simulations different functional depths when there are outliers in the sample. Among the notions of depth we compare, there is one that satisfies the properties of statistical functional depth. This is the one that results in a better performance under the presence of outliers.

CO021 Room Triton 1+2 CLUSTERING AND DIMENSIONALITY REDUCTION Chair: Maria Brigida Ferraro

C0214: **Model-based biclustering of multivariate longitudinal trajectories**

Presenter: **Maria Francesca Marino**, University of Florence, Italy
Co-authors: Marco Alfo, Francesca Martella

Model-based clustering represents nowadays a popular tool of analysis thanks to its probabilistic foundations and its great flexibility. To deal with multivariate longitudinal sequences, standard approaches need to be extended to accommodate the peculiarities of such kind of data. These come in the form of three-way data: the first dimension identifies individuals, the second dimension identifies variables, the third one identifies time occasions. A method for simultaneous clustering of subjects and multivariate outcomes repeatedly recorded over time is proposed. In particular, a finite mixture of generalized linear models is considered to cluster individuals; within each component of the finite mixture, a flexible and parsimonious parameterization of the corresponding canonical parameter is adopted to identify clusters of outcomes evolving in a similar manner across time. This allows us to obtain clusters of individuals that share common trajectories for one of more outcomes over time and, consequently, a dimensionality reduction on the first two dimensions of three-way data structure. Parameter estimates are derived within a maximum likelihood framework, by considering an indirect approach based on an extended expectation-maximization algorithm.

C0199: **A distance for time series with applications to clustering and classification**

Presenter: **Pablo Montero-Manso**, Monash University, Australia
Co-authors: Jose Vilar

Time series are ubiquitous data objects, yet highly complex due to temporal dependence, high dimensionality and often needed invariances such as phase invariance. This complexity often causes standard multivariate techniques to achieve poor results. We propose a new distance between time series based on comparing autoregressive distributions in a nonparametric way. The autoregressive structures are captured via lag embeddings, which are then compared using a divergence measure between multivariate sets originally designed as a multivariate statistic for two-sample testing. The proposed approach achieves very good results in clustering and classification tasks, in both real and simulated data, outperforming or being highly competitive when compared with other distances specifically designed to deal with time series. Furthermore, our approach retains a high level of simplicity and interpretability that may be desired in some contexts, thus opening the possibility of application to general purpose distance-based data analysis methods such as visualization, database search or anomaly detection.

C0204: **Topological low dimensional learning of high dimensional time series**

Presenter: **Tullia Padellini**, Sapienza University of Rome, Italy
Co-authors: Pierpaolo Brutti

As the complexity and the dimension of available data increases, so does the need to characterize them through lower dimensional structures. Topological features are gaining momentum in this quest for insights on the data, as they provide an interpretable description of the connectivity structure of data. We introduce a new topological statistic, the Persistence Flamelet, tailored for high dimensional time series, where we need to summarize data at each time point, as well as their evolution in time. The proposal allows us for visualization of the evolution of hidden structures in the data, and also provides a quantitative measure of their relevance in explaining the data (persistence). After assessing its theoretical properties, such as convergence and stability, we show the performance of this new tool in visualisation as well as in inferential challenges. Focusing on EEG data, whose dependency structure is especially complicated due to the unclear spatial propagation of the signal, we show how topological information can be exploited to explain and recover groups in the observed subjects.

C0218: **Parsimonious and robust hidden semi-Markov models with application to financial time-series**

Presenter: **Antonello Maruotti**, Libera Universita Maria Ss Assunta, Italy
Co-authors: Antonio Punzo, Jan Bulla

A new class of parsimonious models is developed to perform time-varying clustering and dimensionality reduction for the analysis of stock market returns, while accounting for atypical observations. The problem of similarity search in time-series data is addressed by specifying a high-dimensional hidden semi-Markov model. The marginal distribution of returns is described by a mixture of multivariate contaminated-normal (CN) distributions. Compared to the normal distribution, the CN has an additional parameter accounting for the presence of atypical observations, and this allows us a better fit to both the distributional and dynamic properties of daily returns. For the maximum likelihood estimation of the model parameters, we outline an ad-hoc Alternating Expected Conditional Maximization (AECM) algorithm. As an illustration, we provide an example based on the analysis of a Standard and Poor's 500 time series.

Tuesday 16.04.2019 11:50 - 13:00 Parallel Session N – CRONOSMDA2019

CI005 Room Triton 1+2 CRONOS SESSION IV Chair: Tim Verdonck

C0180: **Prediction and robustness**
Presenter: **Elvezio Ronchetti**, University of Geneva, Switzerland
Co-authors: Setareh Ranjbar, Stefan Sperlich

A general discussion of the robustness issues in a prediction framework is provided, and their implications in different areas, including classification, insurance, and estimation in finite populations are analyzed. We illustrate more specifically these issues in the prediction of nonlinear indices (such as inequality or poverty measures) for small areas and in the presence of outliers. We propose two approaches to calibrate for the bias of nonlinear functionals, such as the Gini index and when the so called representative outliers come from a skewed heavy tail distribution.

C0244: **Informative transformation of responses that can be positive or negative**
Presenter: **Marco Riani**, University of Parma, Italy
Co-authors: Anthony Atkinson, Aldo Corbellini

The parametric family of power transformations to approximate normality analysed by Box and Cox can be applied only to positive data. This transformation has been generalized to allow for the inclusion of zero and negative response values, which arise for example in data on GNP growth and company profits and in the differences in measurements before and after treatment. The aim is to describe the use of constructed variables to provide an approximate score statistic for the transformation which avoids the numerical optimization required for estimation of the transformation parameter using maximum likelihood. The resulting statistic is based on aggregate properties of the data. Robust analysis of the data with the forward search provides a series of subsets of the data of increasing size, ordered by closeness to the fitted model for each subset size. The "fan plot" of the statistics for these subsets against subset size clearly indicates the effect of individual observations, especially outliers, on the estimated transformation parameter. The score test is extended to determine whether positive or negative observations require different transformations, leading to an informative extended fan plot. The methods will be illustrated with several examples.

C0224: **Blind source separation based on robust autocovariance matrices**
Presenter: **Sara Taskinen**, University of Jyvaskyla, Finland
Co-authors: Jari Miettinen, Klaus Nordhausen, David Tyler

Assume that the observed p time series are linear combinations of p latent uncorrelated weakly stationary time series. The aim in blind source separation (BSS) is then to find an estimate for the unmixing matrix which transforms the observed time series back to uncorrelated latent time series. In the classical SOBI (Second Order Blind Identification) method, approximate joint diagonalization of the sample covariance matrix and sample autocovariance matrices with several lags are used to estimate the unmixing matrix. However, it is well known that in the presence of outliers, the sample covariance matrix and sample autocovariance matrices perform poorly and yield to unreliable unmixing matrix estimates. We use so-called M-autocovariance matrices in the BSS estimation. M-autocovariance matrices are similar to the classical M-estimators in that they downweight the outliers using some preselected, bounded weight function. We use finite-simple simulation studies and a real data example to illustrate the performance of the robust SOBI method.

CI098 Room Triton 3 SPRING COURSE SESSION XI Chair: Zlatko Drmac

C0280: **Numerical linear algebra for computational statistics**
Presenter: **Zlatko Drmac**, University of Zagreb, Croatia

Why has a covariance matrix negative eigenvalues and other questions related to numerical procedures in computational statistics are discussed, in particular on eigenvalues, singular value decomposition (SVD) and its generalization, the GSVD. These are the tools of trade in various applications, including computational statistics, least squares modeling, vibration analysis in structural engineering - just to name a few. In essence, the GSVD can be reduced to the SVD of certain products and quotients of matrices. For instance, in the canonical correlation analysis of two sets of variables x, y, with joint distribution and the covariance matrix $C = (C_{xx}, C_{xy}; C_{yx}, C_{yy})$, wanted is the SVD of the product $C_{xx}^{-1/2} C_{xy} C_{yy}^{-1/2}$. However, using software implementations of numerical algorithms is not that simple, despite availability of many well-known state-of-the-art software packages. We will review the recent advances in this important part of numerical linear algebra, with particular attention to *(i)* understanding the sensitivity and condition numbers; *(ii)* numerical robustness and limitations of numerical algorithms; *(iii)* careful selection and deployment of reliable mathematical software to be able to interpret and use the computed output with confidence in concrete applications. We illustrate the theoretical numerical issues on selected tasks from computational statistics.

CO076 Room Neptune MULTIVARIATE AND FUNCTIONAL DATA ANALYSIS Chair: Alicia Nieto-Reyes

C0245: **Numerical strategies for the estimation of functional regression models**
Presenter: **Stella Hadjiantoni**, University of Kent, UK, United Kingdom
Co-authors: Ana Colubi, Erricos John Kontoghiorghes

In functional data analysis, the discrete observed data are converted to smooth functions and so they become infinite dimensional data objects. The analysis involves representing the functional data using a basis expansion and then truncating the basis in term of a finite number of basis elements. Choosing the number of basis elements is part of the data analysis. Therefore, the dimension of the basis expansion is an unknown parameter and investigation is required to determine its value. A recursive numerical method is examined for choosing the number of basis elements within the context of model selection. Penalised least squares and cross validation procedures are used in order to choose the number of basis elements that optimise the estimation of the functional regression model. The proposed numerical method is based on orthogonal and hyperbolic transformations.

C0255: **Investigating multimodality in a multivariate setting**
Presenter: **Jose Ameijeiras-Alonso**, KU Leuven, Belgium
Co-authors: Rosa Crujeiras, Alberto Rodriguez-Casal

The identification of the number of modes in a multivariate setting has become an important problem in many applied fields. Complex measurements exhibit some characteristics that cannot be reflected by unimodal densities. Also, classic techniques for identifying the number of subgroups in a population make use of some assumptions that may not be satisfied by the data. For that reason, different approaches were introduced from a nonparametric perspective with the objective of determining the number of modes or their estimated locations. With the aim of providing a way of nonparametrically testing the hypothesis of unimodality in the multivariate setting, a new proposal will be introduced. The correct behavior of the test and the finite-sample properties are studied asymptotically and in a simulation experiment.

C0268: **Statistical local depth for functional data**

Presenter: **Claudio Agostinelli**, University of Trento, Italy

Data depth proves successful in the analysis of multivariate data sets, in particular deriving an overall center and assigning ranks to the observed units. Two key features are: the directions of the ordering, from the center towards the outside, and the recognition of a unique center irrespective of the distribution being unimodal or multi modal. This behavior is a consequence of the ray monotonicity. Recently, a new framework allowing identification of partial centers was suggested. The corresponding generalized depth functions, called local depths are able to record local features and can be used in mode detection, identification of subgroups and in cluster analysis. Functional data are become common nowadays. The half-region depth suited for functional data and for high dimensional data has been proposed. A local half-region depth is introduced and its properties discussed. Several examples will illustrate the use of this new tool in data analysis.

CC065 **Room Business center** FUNCTIONAL AND MULTIVARIATE DATA ANALYSIS **Chair: Enea Bongiorno**

C0275: **Sparsely observed functional time series: Estimation and prediction**

Presenter: **Tomas Rubin**, EPFL, Switzerland

Co-authors: Victor Panaretos

Functional time series analysis has traditionally been carried out under the assumption of complete observation of the constituent series of curves, assumed stationary. Nevertheless, it may well happen that the data available to the analyst are not the actual sequence of curves, but relatively few and noisy measurements per curve, potentially at different locations in each curve's domain. The subject is to tackle the problem of estimating the dynamics and of recovering the latent process of smooth curves in this sparse observation regime. We construct a consistent nonparametric estimator of the series' spectral density operator and use it develop a frequency-domain recovery approach, that predicts the latent curves at a given time by borrowing strength from the (estimated) dynamic correlations in the series across time. Further to predicting the latent curves from their noisy point samples, the method fills in gaps in the sequence (curves nowhere sampled), denoises the data, and serves as a basis for forecasting. Means of providing corresponding confidence bands are also investigated. The methodology is further illustrated by application to an environmental data set on fair-weather atmospheric electricity, which naturally leads to a sparse functional time-series.

C0281: **Dynamic modelling of functional data using warped solutions of SDEs**

Presenter: **Niels Lundtorp Olsen**, University of Copenhagen, Denmark

Co-authors: Anders Tolver

Many models for functional data assume that temporal and amplitude variation can be separated. However, complex interactions may exist e.g. when data are obtained from biomechanical systems where changing the frequency will likely influence the amplitude of a signal. We suggest to jointly model the phase and amplitude variation of a function by noisy differential equations, i.e. stochastic differential equations, but crucially the SDEs are inhomogeneous with a warped time argument.We restrict ourselves to second-order SDEs on the form: $\dot{X}(t) = F(X(t), \dot{X}(t), v(t))dt + \sigma dW_t$. Based on discrete and noisy observations of the signals solving the SDE we demonstrate how maximum likelihood estimation may be carried out. Under certain linearity conditions, the Kalman filter can be applied, and used together with MCMC methods for estimation. We demonstrate the method on simulated and real data. One important example is functional data with locally periodic structures arising from applying periodic SDEs with warped time arguments. The approach allows us to model the influence of the local internal speed of the signal on the dynamical relation between derivatives. In particular, it has the potential of predicting the functional response when altering the timing of the system.

C0292: **Detecting clusters in the data from variance decompositions of its projections**

Presenter: **Yannis Yatracos**, Cyprus University of Technology, Cyprus

A new projection-pursuit index is used to identify clusters and other structures in multivariate data. It is obtained from an unusual variance decomposition of the data's one-dimensional projections, without assuming a model for the data or that the number of clusters is known. The index is affine invariant and successful with real and simulated data. A general result is obtained indicating that clusters' separation increases with the data's dimension. In simulations it is thus confirmed, as expected, that the performance of the index either improves or does not deteriorate when the data's dimension increases, making it especially useful for "large dimension-small sample size" data. The efficiency of this index will increase with the continuously improved computer technology and additional statistical analysis for each problem. Applications and open problems are presented.

Tuesday 16.04.2019 14:30 - 16:00 Parallel Session O – CRONOSMDA2019

CI100 Room Triton 3 SPRING COURSE SESSION XII Chair: Daniela Zaharie

C0261: **Machine learning methods for multivariate data analysis**
Presenter: **Daniela Zaharie**, West University of Timisoara, Romania

Predictive tasks (e.g. classification or regression) can be solved by using various machine learning models (e.g. k-nearest neighbours, decision trees, support vector machines, neural networks etc). The prediction accuracy of individual models can be improved by aggregating several models using various ensemble techniques (e.g bagging, boosting, stacking). Besides these explicit ensemble techniques, there are also strategies (e.g. dropout) which induce an implicit ensemble with shared parameters by injecting extra randomness into the machine learning model and therefore generating various model instances which are then aggregated. On the other hand, in real-world applications it would be useful to provide a measure for the prediction uncertainty. Most of the machine learning models, particularly the black-box ones (e.g. neural networks), do not provide directly estimates of the prediction uncertainty. However the information provided by ensemble models can be exploited in order to estimate uncertainty measures. The aim is to provide an overview on meta-models with a focus on ensemble strategies applied to decision trees (e.g. random forests, boosted decision trees). The particularities of randomly dropping out parameters of the model and its impact on the performance are discussed. Several approaches in estimating the uncertainty of the prediction are discussed in the context of solving predictive tasks in biology and for semantic segmentation of satellite images.

CO050 Room Neptune DATA DEPTH AND MULTIVARIATE QUANTILES Chair: Stanislav Nagy

C0240: **Empirical and population level sets of Tukey's halfspace depth**
Presenter: **Victor-Emmanuel Brunel**, ENSAE ParisTech, France

Tukey's halfspace depth has attracted much interest in data analysis, because it is a natural way of measuring the notion of depth relative to a cloud of points or, more generally, to a probability measure. Given an i.i.d. sample, we investigate the concentration of upper level sets of the Tukey depth relative to that sample around their population version. We show that under some mild assumptions on the underlying probability measure, concentration occurs at a parametric rate and we deduce moment inequalities at that same rate. In a computational prospective, we also study the concentration of a randomly discretized version of the empirical upper level sets of Tukey depth.

C0212: **Weighted halfspace depth**
Presenter: **Daniel Hlubinka**, Univerzita Karlova, Matematicko-fyzikalni fakulta, Czech Republic
Co-authors: Lukas Kotik

Statistical depth functions are well-known nonparametric tools for analyzing multivariate data. Halfspace depth is most frequently used, and while it has many desirable properties, it is dependent on global characteristics of the underlying distribution. In some circumstances, however, it may be desirable to take into account more local and intrinsic characteristics of the data. To this end, we introduce weighted halfspace depths in which the indicator function of closed halfspace is replaced by a more general weight function. Our approach, which calls in part on functions associated with conic sections, encompasses as special cases the notions of sample halfspace depth and kernel density estimation. We give several illustrations and prove the strong uniform consistency of weighted halfspace depth incorporating mild conditions on the weight function.

C0186: **Tests for validity of the semiparametric heteroskedastic transformation model**
Presenter: **Charl Pretorius**, Charles University, Czech Republic
Co-authors: Simos Meintanis, Marie Huskova

There exists a number of tests for assessing the nonparametric heteroscedastic location-scale assumption. We consider a goodness-of-fit test for the more general hypothesis of the validity of this model under a parametric functional transformation on the response variable. Specifically we consider testing for independence between the regressors and the errors in a model where the transformed response is just a location/scale shift of the error. Our criteria use the familiar factorization property of the joint characteristic function of the covariates under independence. The difficulty is that the errors are unobserved and hence one needs to employ properly estimated residuals in their place. We provide the limit distribution of the test statistics under the null hypothesis as well as under alternatives, and also suggest a resampling procedure in order to approximate the critical values of the tests. This resampling is subsequently employed in a series of Monte Carlo experiments that illustrate the finite-sample properties of the new test. We also investigate the performance of related test statistics for normality and symmetry of errors.

C0174: **Statistical depth functions in robust causality analysis of economic phenomena**
Presenter: **Daniel Kosiorowski**, Cracow University of Economics, Poland
Co-authors: Dominik Mielczarek, Jerzy Rydlewski

Often the perception of an economic phenomenon relates to an evaluation of properties of a function of a certain continuum. One may consider GDP per capita trajectory of a country during a decade, day and night number of visits on an internet service of a user, or a behaviour of an investor optimism indicator within a month. Reducing the whole function to a certain set of scalars (e.g., mean, variance) very often denotes a significant loss of valuable information on the phenomenon and in a consequence may lead to inappropriate perception of the phenomenon. A "shape" of the consumer price index (CPI) during a month may better than a set of descriptive statistics express an investor optimism during the considered period as a specific sequence of peaks and valleys in CPI trajectory may denote sequence of an activity bursts and consumer hesitations hence "a spectrum of moods" called the optimism. Causality tells us not only that two phenomena are related, but how they are related. It allows us to make robust prediction about the future, explains the relationships between and occurrence of events and enables to develop effective policies intervention. We critically discuss certain chances, dangers and challenges related to an application of tools offered by the so called data depth concept for functional data in "robust causality analysis of economic phenomena'. Theoretical considerations are illustrated by means of real economic examples.

CO074 Room Triton 1+2 OPTIMAL TRANSPORT METHODS IN STATISTICS Chair: Yoav Zemel

C0191: **Empirical optimal transport: Inference, algorithms, applications**
Presenter: **Axel Munk**, Georg-August-University Goettingen, Germany

Recent developments in statistical data analysis based on empirical optimal transport (EOT) are discussed. Fundamental are limit laws for EOT plans and distances on finite and discrete spaces. These are characterized by dual optimal transport problems over a Gaussian process. Our proofs are based on a combination of sensitivity analysis from convex optimization and discrete empirical process theory. We examine an upper bound for such limiting distributions based on a spanning tree approximation which can be computed explicitly. This can be used for statistical inference, fast simulation, and for fast randomized computation of optimal transport in large scale data applications at specified computational cost as it provides bounds to balance computational and statistical error. Our methodology is illustrated in computer experiments and on biological data from super-resolution cell microscopy. Finally, this is contrasted and compared with recent results on regularized empirical optimal transport.

C0286: **Synthesizing facial photometries and corresponding geometries using generative adversarial networks**
Presenter: **Gil Shamai**, Technion, Israel
Co-authors: Ron Slossberg, Ron kimmel

Artificial data synthesis is currently a well studied topic with useful applications in data science, computer vision, graphics and many other fields. Generating realistic data is especially challenging since human perception is highly sensitive to non realistic appearance. In recent times, new levels of realism have been achieved by advances in GAN training procedures and architectures. These successful models, however, are tuned mostly for use with regularly sampled data such as images, audio and video. We propose a new method for generating realistic human facial geometries coupled with overlayed textures. We circumvent the parametrization issue by imposing a global mapping from our data to the unit rectangle. We address the often neglected topic of relation between texture and geometry and propose to use this correlation to match between generated textures and their corresponding geometries. We offer a new method for training GAN models on partially corrupted data. Finally, we provide empirical evidence demonstrating our generative model's ability to produce examples of new identities independent from the training data while maintaining a high level of realism, two traits that are often at odds.

C0194: **Bayesian learning with Wasserstein barycenters**
Presenter: **Julio Backhoff-Veraguas**, University of Vienna, Austria
Co-authors: Joaquin Fontbona, Gonzalo Rios, Felipe Tobar

A novel paradigm for Bayesian learning is introduced based on optimal transport theory. Namely, we propose to use the Wasserstein barycenter of the posterior law on models as a predictive posterior, thus introducing an alternative to classical choices like the maximum a posteriori estimator and the Bayesian model average. We exhibit conditions granting the existence and statistical consistency of this estimator, discuss some of its basic and specific properties, and provide insight into its theoretical advantages. Finally, we introduce a novel numerical method which is ideally suited for the computation of our estimator. This method can be seen as a stochastic gradient descent algorithm in the Wasserstein space.

C0290: **Convergence rates for empirical barycenters in metric spaces**
Presenter: **Quentin Paris**, NRU HSE, Faculty of computer science, Russia

Rates of convergence are presented for empirical barycenters of a probability measure on a metric space under general conditions. The results connect ideas from metric geometry to the theory of empirical processes and is studied in two meaningful scenarios. The first one is a geometrical constraint on the underlying space referred to as k-convexity, compatible with a positive upper curvature bound in the sense of Alexandrov. The second scenario considers the case of a non-negatively curved space on which geodesics, emanating from a barycenter, can be extended. While not restricted to this setting, our results are discussed in the context of the Wasserstein spaces.

CC064 Room Business center MULTIVARIATE AND APPLIED DATA ANALYSIS Chair: Andreas Artemiou

C0271: **On modified interdirections and lift-interdirections**
Presenter: **Jana Klicnarova**, The Czech Academy of Sciences, Czech Republic

Multivariate signs and ranks give rise to many nonparametric statistical procedures. Those based on hyperplane-based concepts of interdirections and lift-interdirections are very appealing due to their affine invariance and clear geometric interpretation, but they also have a few disadvantages including intractable computation in high dimensions. It turns out that the concepts can be modified to preserve all the good properties, overcome certain drawbacks and still be applicable in many sign and rank tests without any loss of efficiency. The research was supported by the Czech Science Foundation project GA17-07384S.

C0274: **Compositional analysis of untargeted metabolomic data using multiple Bayesian hypotheses testing**
Presenter: **Julie Rendlova**, Palacky University, Czech Republic
Co-authors: Karel Hron, Ondrej Vencalek, David Friedecky, Tomas Adam

Currently, both clinical targeted and untargeted metabolomic approaches aim to find statistically significant differences in chemical fingerprints of patients with some disease and a control group and to identify biological markers allowing a prediction of the disease. Traditionally, the differences between controls and patients are evaluated by both univariate and multivariate statistical methods. The univariate approach relies merely on t-tests (or their nonparametric version) where the results from multiple testing are compared by p-values and fold-changes using a so-called volcano plot. As a counterpart, a multiple Bayesian hypotheses testing is proposed, introducing a concept of b-values as well as a Bayesian version of the volcano plot incorporating distance levels of the posterior highest density intervals from zero. Moreover, since each metabolome is a collection of some small-molecule metabolites in a biological material, relative structure of metabolomic data is of the main interest. A proper choice of orthonormal coordinates w.r.t. Aitchison geometry considering the compositional character of a metabolome is, therefore, an essential step in any statistical analysis of such data. The theoretical background is accompanied by an analysis of a data set containing dry blood spots of patients suffering from an inherited metabolic disorder in beta oxidation of fatty acid metabolism - medium-chain acyl-CoA dehydrogenase deficiency (MCADD).

C0283: **Robust sparse generalized dynamic PCA: Algorithms and some applications**
Presenter: **Jan Bruha**, CNB, Czech Republic

A simple extension of the Generalized Dynamic PCA is introduced. We allow for (1) sparsity, (2) robustness and discuss (3) the possibility of limited time variation. Originally, the robustness was achieved by using robust M-estimator. We instead consider trimmed-least squares criterion that can handle a larger class of outliers. The sparsity is achieved by restricting some of the loadings to zero. The goal is to increase statistical efficiency and to allow for a more straightforward interpretation of principal components. We provide a practical numerical algorithm.

C0152: **Mixed effect parameters 4-variate Gompertz type diffusion process: Computational aspects and information measures**
Presenter: **Petras Rupsys**, Aleksandras Stulginskis University, Lithuania

A 4-variate diffusion process is used to parameterize tree size variables data in forest stands. A diffusion process is defined by a mixed effect parameters 4-variate stochastic differential equation (SDE). We propose a system of the 4-variate mixed effect parameters Gompertz type SDE to quantify the dynamics via stand age of the probability density function of the tree size variables with a sigmoid form trend for the mean values. The SDE model allows us a better understanding of biological processes driving the dynamics of natural phenomena. The new derived 4-variate probability density function and its marginal univariate, bivariate and trivariate, and conditional univariate, bivariate and trivariate distributions can be applied for the modeling of stand attributes such as the mean diameter, height, stem and stand volumes. All parameters are estimated by the maximum likelihood procedure using real-life dataset. We have introduced a general mutual information to capture multivariate interactions between tree size variables. We have experimentally confirmed that 4-variate mutual information quantities are appropriate measures to reconstruct multivariate interactions in tree size variables. The results are implemented in the symbolic algebra system MAPLE.

Tuesday 16.04.2019 16:30 - 18:30 Parallel Session P – CRONOSMDA2019

CI102 Room Triton 3 SPRING COURSE SESSION XIII Chair: Ivette Gomes

C0156: **Statistics of extremes and risk assessment using R**
Presenter: **Ivette Gomes**, FCiencias.ID, Universidade de Lisboa and CEAUL, Portugal

Extreme value theory (EVT) helps us to control potentially disastrous events, of high relevance to society and with a high social impact. Floods, fires, and other extreme events have provided impetus for recent re-developments of EV analysis. In EVT, just as in almost all areas of statistics, the ordering of a sample is of primordial relevance. After a brief reference to a few concepts related to ordering, we provide some motivation for the need of EVT in the analysis of rare events, in fields as diverse as environment, finance and insurance, among others. Next, the general EV and the generalized Pareto distributions are introduced, together with the concepts of extreme value index and the notion of tail-heaviness. Finally, we deal with several topics in the field of statistics of extremes, an highly useful area in applications, whenever we want to make inference on tails, estimating rare events parameters, either univariate or multivariate. Apart from providing a review of most of the parametric approaches in the area, we further refer a few semi-parametric approaches, with the analysis of case-studies in the aforementioned fields, performed essentially through the use of R-packages for extreme values, like the evd, evdbayes, evir, ismev, extRemes, fExtremes, POT, and SpatialExtremes, among others.

CO106 Room Triton 1+2 INTER-COST ACTIONS: APPLIED DATA ANALYSIS Chair: Ana Colubi

C0296: **Nowcasting the public's response to natural hazards in the google era**
Presenter: **Konstantinos Tsagarakis**, Democritus University of Thrace, Greece
Co-authors: Amaryllis Mavragani

In the Big Data era, employing online sources to analyze and predict human behavior is significantly increasing. Internet data have the advantage of providing the revealed instead of the stated users' preferences, thus giving us access to information that would not be accessible otherwise. Online search traffic data are suggested to be valuable in forecasting, mostly in the fields of health, the environment, and economics. Towards this direction, the aim is to examine the online behavior and reaction towards natural phenomena, e.g., forest fires, floods, and tornados. Next, the correlations between online queries and the respective indices are estimated, in order to explore the possibility of nowcasting the public's behavior. Preliminary analysis northern hemisphere countries shows that the public's reaction towards such natural hazards is immediately depicted on online queries. Seasonality of the online activity follows that of the hazards, as in the case of forest fires, peaking during the summer months. The present analysis has significant policy implications, as retrieving data from online sources tackles the issues caused by the traditional data collection methods, especially in developing social indices that require real time data for immediate and robust results. Overall, in finding new, innovative ways of retrieving real time data, online query data have been shown to be of much value, as they correlate with real life data and are undoubtedly valuable in forecasting.

C0293: **Assessing an increase of extreme rainfall by a bootstrap isotonic test**
Presenter: **Elena Fernandez Iglesias**, University of Oviedo, Spain
Co-authors: Gil Gonzalez-Rodriguez, Jorge Marquinez

Rainfall storm events has been identified by considering a 0.995 quantile in the hourly precipitation series of six meteorological stations located in the NW of Spain. The longer period under analysis covers 46 years. Taking into account previous studies developed in this region as well as the results of the wide-sense isotonic Bootstrap tests a non-decreasing trend for most of the series can be assumed. That is, the regression functions under analysis could be constant or otherwise they present an effective increment at certain (or several) time points. In order to assess an effective increment of the storm series a Bootstrap isotonic contrast in strict sense was applied. Although there is an important variability, the obtained results identify an effective increment in expected total rainfall, total time and the number of storm events per year in some series at a significance level of 5%. Besides, the results indicates an increase in the mean rainfall of the storm events from the West to the East of the Cantabrian region. The behaviour identified is consistent with the forecast global climate change in this part of Europe and with an increase of extreme rainfall.

C0294: **Deep learning for biological sequence analysis: The SignalP software**
Presenter: **Konstantinos Tsirigos**, Technical University of Denmark, Denmark

Deep learning models have been successful in numerous applications such as image, text, and speech recognition. They have become increasingly popular amongst the machine learning tools for bioinformatics during the recent years owing to the availability of greater computational resources, more data, new training algorithms and easy-to-use libraries for implementation. We have used deep learning in a widely-known topic of biological protein sequence analysis and, specifically, the detection of signal peptides in amino-acid sequences. Signal peptides are intrinsic signals for secretion in both eukaryotic and prokaryotic proteins. Since their existence was demonstrated in 1975 by Gnter Blobel (who later received the Nobel Prize for it), there has been a keen interest in the question of how signal peptides actually look and whether they can be predicted from the amino-acid sequence alone. One of the most used methods for making such predictions, SignalP, which has been online since 1996, is now released in its fifth major version, using a deep learning architecture.

C0295: **Using data analysis to learn from an infrastructure dataset for highway bridges**
Presenter: **Dimos Charmpis**, University of Cyprus, Cyprus
Co-authors: Filippos Alogdianakis, Ioannis Balafas

Various infrastructure information is gathered nowadays in databases, which have become rather large after years of development and data collection. For thorough search and broad exploitation of the available information, even beyond its original scope, data analysis approaches need to be employed. The aim is to exploit the data in the US National Bridge Inventory (NBI) maintained by the Federal Highway Administration (FHWA), which includes information for over 500,000 bridges. The information provided in NBI was analysed in combination with additional data from other sources (for climatic conditions, earthquake hazard, etc.). Where needed, data were converted to correspond to bridge locations using spatial interpolation techniques. Then, Exploratory Data Analysis (EDA), Analysis of Variance (ANOVA) and regression analysis methods were utilized to study the causes of bridge deterioration. These statistical methods yield quantitative results and allow the identification, ranking and measurement of intensity of factors contributing to the decrease of the structural condition of bridges with time.

C0285: **A semiparametric default forecasting model**
Presenter: **Christakis Charalambous**, University of Cyprus, Cyprus
Co-authors: Spiros Martzoukos, Zenon Taoushianis

A fundamental limitation of structural models for the estimation of the probability of default is that their most important parameters, the value of assets and volatility, are not observed in the market. We develop a methodology where the unobserved parameters are viewed as generalized functions. Using a nonparametric approach for their estimation, we obtain improved parameter values which enter the parametric model, yielding a semi-parametric model. In this context, the Black-Scholes-Merton model is used as a paradigm. Results show substantial improvement in the out-of-sample performance when comparing our semiparametric model with other alternative specifications of the Black-Scholes-Merton model in terms of discriminatory power, information content and economic impact.

Authors Index