Book of Abstracts

COMPSTAT 2016

22nd International Conference on Computational Statistics

August 23-26, 2016

Satellite CRoNoS Workshop and Summer Course Functional Data Analysis

August 26-28, 2016

August 23-28, 2016

Auditorio Príncipe Felipe, Oviedo, Spain

PROGRAMME AND ABSTRACTS

22nd International Conference on

Computational Statistics (COMPSTAT 2016)

http://www.compstat2016.org

Auditorium/Congress Palace Principe Felipe, Oviedo, Spain
23-26 August 2016

2016 CRoNoS Summer Course and Satellite Workshop on

Functional Data Analysis (CRoNoS FDA 2016)

http://www.compstat2016.org/CRoNoS_SummerCourse.php
http://www.compstat2016.org/SatelliteWorkshop.php

Auditorium/Congress Palace Principe Felipe, Oviedo, Spain
26-28 August 2016

Technical Editors: Angela Blanco-Fernandez and Gil Gonzalez-Rodriguez.

COMPSTAT 2016 Scientific Program Committee:

Ex-officio:

COMPSTAT 2016 organiser and Chairperson of the SPC: Ana Colubi.
Past COMPSTAT organiser: Manfred Gilli.
Next COMPSTAT organiser: Cristian Gatu.
Incoming IASC-ERS Chairman: Alessandra Amendola.

Members:

Marc Genton, Salvatore Ingrassia, Jean-Michel Poggi, Igor Pruenster, Juan Romo, Tamas Rudas and Stefan Van Aelst.

Consultative Members:

Representative of the IFCS: Christian Hennig.
Representative of the ARS of IASC: Chun-houh Chen.
Representative of CMStatistics: Erricos Kontoghiorghes.

COMPSTAT2016 Proceedings Management Committee:

Ana Colubi, Angela Blanco and Cristian Gatu.

Local Organizing Committee:

Ana M. Aguilera, Gil Gonzalez-Rodriguez, M. Dolores Jiménez-Gamero, Agustin Mayo, Domingo Morales and M. Carmen Pardo.

CRoNoS FDA 2016 Scientific Program Committee:

Ana M. Aguilera, Ana Arribas-Gil, Enea Bongiorno, Frederic Ferraty, Wenceslao González-Manteiga, Alois Kneip and Juan Romo.

Organizers:

Ana Colubi and Gil Gonzalez-Rodriguez.

Dear Friends and Colleagues,

We wish to warmly welcome you to Oviedo, for the 22nd International Conference on Computational Statistics (COMPSTAT 2016) and the CRoNoS Summer Course and Satellite Workshop on Functional Data Analysis (CRoNoS FDA 2016). These events are locally organized by members of the University of Oviedo assisted by active Spanish researchers. The COMPSTAT is an initiative of the European Regional Section of the International Association for Statistical Computing (IASC-ERS), a society of the International Statistical Institute (ISI). COMPSTAT is one of the best-known world conferences in Computational Statistics, regularly attracting hundreds of researchers and practitioners.

The first COMPSTAT conference took place in Vienna in 1974, and the last two editions took place in Limassol in 2012 and Geneve in 2014. It has gained a reputation as an ideal forum for presenting top quality theoretical and applied work, promoting interdisciplinary research and establishing contacts amongst researchers with common interests.

Keynote lectures are addressed by Prof. Gerard Biau, Université Pierre et Marie Curie, Paris, France, Prof. Alastair Young, Imperial College, London, UK and Prof. Hans-Georg Mueller, University of California Davis, United States.

From more than 450 submissions received for COMPSTAT, 360 have been retained for presentation in the conference. The conference programme has 41 contributed sessions, 8 invited sessions, 3 keynote talks, 30 organized sessions and 3 tutorials. There are approximately 430 participants. The COST Action IC1408 CRoNoS Summer Course and Satellite Workshop have about 90 participants, 45 talks and 10 hours of lectures.

The Proceedings are published in an electronic book comprising 34 papers. The participants can find and electronic copy in a USB stick placed in their conference bags or download it from the conference web page. All the papers submitted have been evaluated through a rigorous peer review process. Those papers that have been accepted for publication in the Proceedings have been evaluated thoroughly by at least 2 referees. This ensures a high quality proceedings volume in the main areas of computational statistics.

The organization would like to thank the editors, authors, referees and all participants of COMPSTAT 2016 who contributed to the success of the conference. Our gratitude to sponsors, scientific programme committee, session organizers, local universities, the city of Oviedo, and many volunteers who have contributed substantially to the conference. We acknowledge their work and support.

The COMPSTAT 2016 organizers invite you to the next edition of the COMPSTAT, which will take place in Iasi, Romania in 2018. We wish the best success to Cristian Gatu the Chairman of the 23rd edition of COMPSTAT.

Ana Colubi
Organiser and Chairperson of the SPC

SCHEDULE

COMPSTAT 2016 - Programme

2016-08-23	2016-08-24	2016-08-25	2016-08-26
Opening, 09:25 - 09:40	**F** 09:00 - 10:30	**I** 09:00 - 10:30	**N** 09:00 - 10:30
A - Keynote 09:40 - 10:30			
Coffee Break 10:30 - 11:00	**Coffee Break** 10:30 - 11:00	**Coffee Break** 10:30 - 11:00	**Coffee Break** 10:30 - 11:00
B 11:00 - 12:50	**G** 11:00 - 12:50	**J** 11:00 - 12:05	**O** 11:00 - 12:05
		K - Keynote 12:15 - 13:05	**P - Keynote** 12:15 - 13:05
Lunch Break 12:50 - 14:30	**Lunch Break** 12:50 - 14:30	**Lunch Break** 13:05 - 14:45	**Closing**, 13:05 - 13:15
C 14:30 - 16:00	**H** 14:30 - 16:00	**L** 14:45 - 16:15	
Coffee Break 16:00 - 16:30		**Coffee Break** 16:15 - 16:45	
D 16:30 - 18:00	**Guided Visit** 16:30 - 18:00	**M** 16:45 - 18:35	**Closing Reception** 18:45 - 20:15
Welcome Reception 20:00 - 21:30	**Concert** 20:00 - 21:30	**Conference Dinner** 20:30 - 23:00	

SCHEDULE

CRoNoS FDA 2016 - Programme

2016-08-26

A - Keynote
12:15 - 13:15

Lunch Break
13:15 - 15:00

B
15:00 - 16:15

Coffee Break
16:15 - 16:45

C
16:45 - 18:25

Welcome Reception
18:45 - 20:15

2016-08-27

D
08:50 - 10:30

Coffee Break
10:30 - 11:00

E - Keynote
11:00 - 11:50

F
12:00 - 13:15

Lunch Break
13:15 - 15:00

G
15:00 - 16:40

Coffee Break
16:40 - 17:10

H
17:10 - 18:50

Workshop and Summer Course Dinner
20:30 - 23:30

2016-08-28

J
08:50 - 10:50

Coffee Break
10:50 - 11:20

K
11:20 - 13:00

Lunch Break
13:00 - 14:30

L
14:30 - 16:10

Coffee Break
16:10 - 16:40

M - Keynote
16:40 - 17:30

TUTORIALS, SUMMER COURSE, MEETINGS AND SOCIAL EVENTS

TUTORIALS - COMPSTAT 2016

The tutorials will take place at room Sala Camara during the conference and in parallel with the invited, organized and contributed sessions. The first is given by Christian Hennig (Cluster validation: How to think and what to do) on Tuesday 23.8.2016, 11:00 - 12:50. The second tutorial is given by Maria Angeles Gil (A methodology to analyze fuzzy data) on Wednesday 24.8.2016, 11:00 - 12:50. Finally, the third tutorial is given by Stephen Pollock (Band pass filtering and wavelets analysis) on Thursday 25.8.2016, 16:45 - 18:35.

SUMMER COURSE - CRoNoS FDA 2016

The summer course will take place at room Sala 1 and in parallel with the organized and contributed sessions of the Satellite CRoNoS Workshop. The course is given by Hans-Georg Mueller and Jane-Ling Wang (Functional Data Analysis: From Basics to Current Topics of Interest) during Friday 26.8.2016 and Saturday 27.8.2016 and by Hannu Oja (Independent component analysis for functional data) during Sunday 28.8.2016.

SPECIAL MEETINGS by invitation to group members

- IASC Executive Committee meeting, *Room: Sala 6*, Tuesday 23rd August 2016, 12:55-14:25.
- ERS BoD Meeting, *Room: Sala 6*, Wednesday 24th August 2016, 12:55-14:25.
- IASC and ERS General Assembly, *Room: Sala Camara*, Thursday 25th August 2016, 18:40-19:40.
- COST Action IC1408 CRoNoS Core Management Group, Lunch meeting, *Room: Sala 6*, Thursday 25th August 2016, 13:05-14:45.

SOCIAL EVENTS - COMPSTAT 2016

- *The coffee breaks* will take place at the Ground Floor Hall. You must have your conference badge in order to attend the coffee breaks.
- *Buffet Lunch* will be arranged at the restaurant on the third floor of the venue on 23rd, 24th, 25th and 26th of August 2016. The lunches are optional and registration is required. The number of places is limited. You must have the corresponding ticket in order to attend the lunch each day. People not registered for lunch can buy lunch at restaurants and cafes in close walking distance to the conference venue.
- *Welcome Reception, Tuesday 23rd of August, 20:00-21:30*. The Welcome Reception will take place at the venue and is open for free to all registrants who had preregistered and accompanying persons who have purchased a reception ticket. Conference registrants must bring their conference badge and any accompanying persons should bring their reception tickets in order to attend the reception. Preregistration is required due to health and safety reasons.
- *Guided Visit, Wednesday 24th of August 2016, 16:30-18:00*. Oviedo is a historical city with a unique architecture. A guided visit for the participants to the conference has been organized for Wednesday 24th of August at 16:30. The meeting point for the guided visit is the Cathedral in the center of the city (see map at page IX). The event is free, but pre-registration is mandatory to obtain your visit voucher.
- *Concert, Wednesday 24th of August 2016, 20:00-21:30.* A concert will take place at the room Sala Principal of venue starting at 20:00. The event is free but pre-registration is mandatory to obtain your concert voucher.
- *Conference Dinner, Thursday 25th August 2016, 21:00-23:30*. The conference dinner is optional and registration is required. It will take place at the Palacio de Exposiciones y Congresos Ciudad de Oviedo (Calatrava Building - see map at page IX). Conference registrants and accompanying persons should bring their conference dinner tickets in order to attend the conference dinner.
- *Closing Reception, Friday 26th of August 2016, 18:45-20:15*. The closing reception will take place at the Hall of the first floor of the venue and is free for all registrants. You must have your conference badge in order to attend the closing reception.

SOCIAL EVENTS - CRoNoS FDA 2016

- *The coffee breaks* will take place at the Ground Floor Hall. You must have your CRoNoS FDA 2016 badge in order to attend the coffee breaks.
- *Welcome Reception, Friday 26th of August, 18:45-20:15*. The Welcome Reception will take place at the venue and is free for all registrants. Summer course and satellite workshop registrants must bring their badge in order to to attend the reception.
- *Summer Course and Satellite Workshop Dinner, Saturday 27th August 2016, 21:00-23:30.* The CRoNoS FDA 2016 dinner is optional and registration is required. It will take place at La Corte restaurant (see map at page IX). CRoNoS FDA 2016 registrants should bring their dinner ticket in order to attend the dinner.

Address of venue

The Conference, Summer Course and Satellite Workshop venue is the Auditorium/Congress Palace Principe Felipe, at Plaza Gesta, 33007 Oviedo, Asturias (Spain).

Registration

The registration will be open from Tuesday 23rd August 2016 and will take place at the Cross Hall of the Auditorium/Congress Palace.

Lecture rooms

The paper presentations will take place at the ground, first and third floors of the Auditorium/Congress Palace. The location of the rooms can be checked at pages X to XII. Rooms Sala de Camara and Sala 7 are in the ground floor, room Sala Principal is in the first floor and Rooms Sala 1-5 are in the third floor. The opening, keynote and closing talks of the COMPSTAT 2016 conference will take place at room Sala de Camara. The poster presentations will take place in the Ground Hall located in the ground floor where the coffee breaks will take place. The first keynote talk of the CRoNoS FDA 2016 Summer Course will take place at room Sala de Camara while the remaining keynote talks will be held in room Sala 1.

Presentation instructions

The lecture rooms will be equipped with a 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 to the session chair with the files for the presentation in PDF (Acrobat) or PPT (Powerpoint) format on a USB memory stick. This must be done ten minutes before each session. The PC in the lecture rooms should be used for presentations. The session chairs are kindly requested to have a laptop for backup. IT technicians will be available during the conference and should be contacted in case of problems. 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

Throughout the venue there will be wireless Internet connection available at the halls. Wifi information will be displayed by the registration desk.

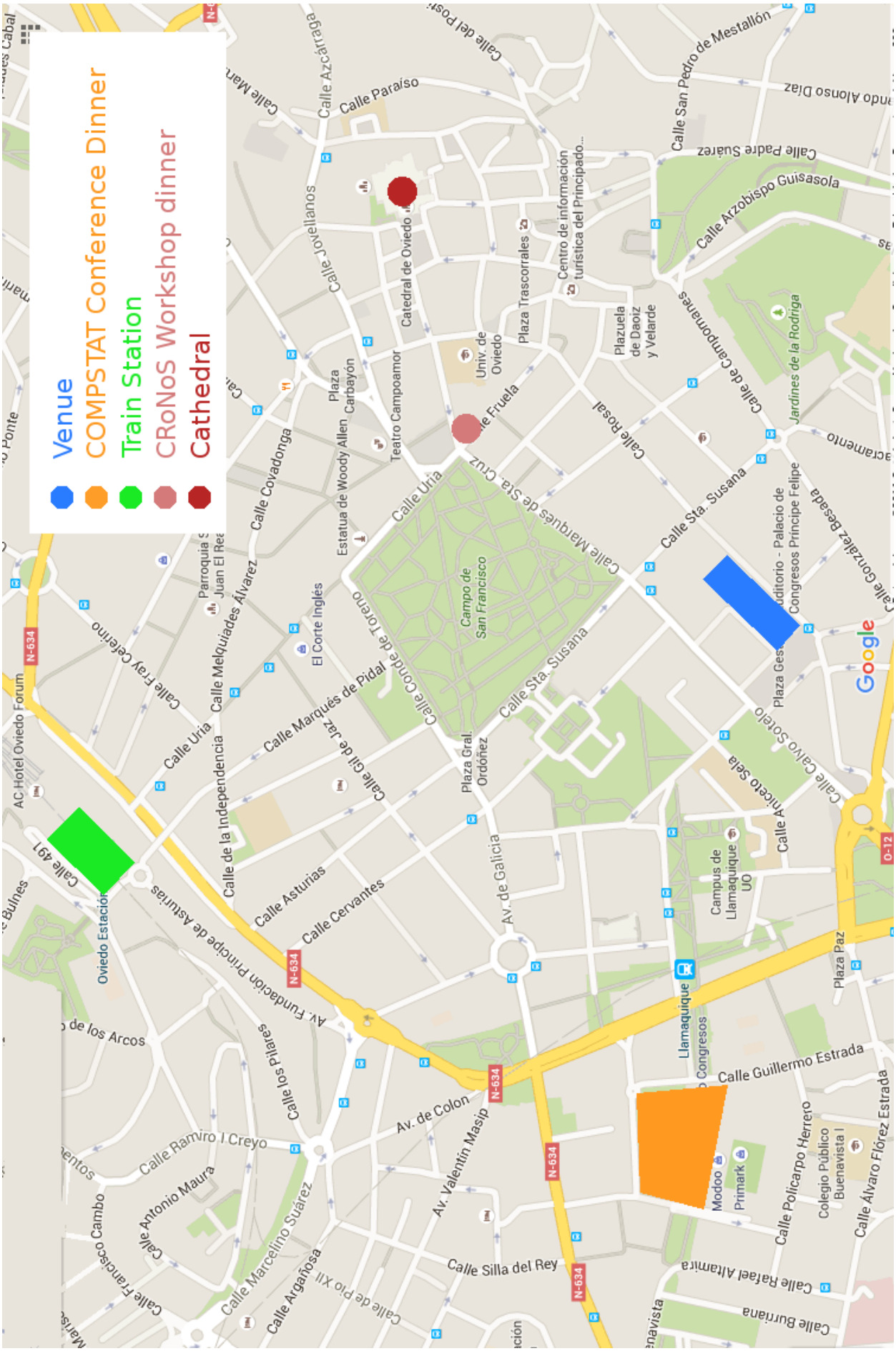
Information and messages

General information about the city, the region, restaurants, useful numbers, etc. can be obtained from the hospitality desk.

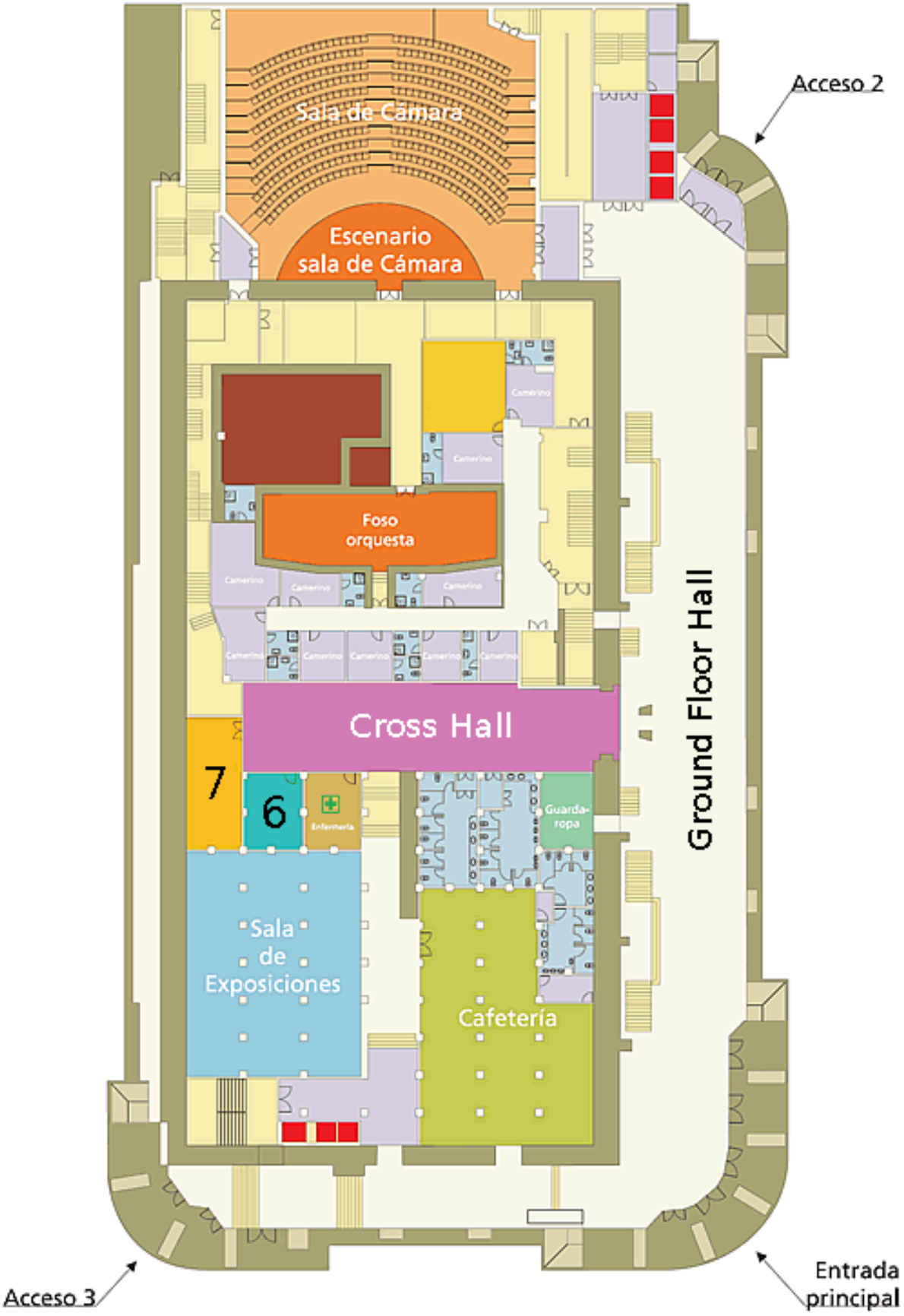
Exhibitors

Elsevier (`http://www.elsevier.com`)
Springer (`http://www.springer.org/`)

Map of the venue and nearby area

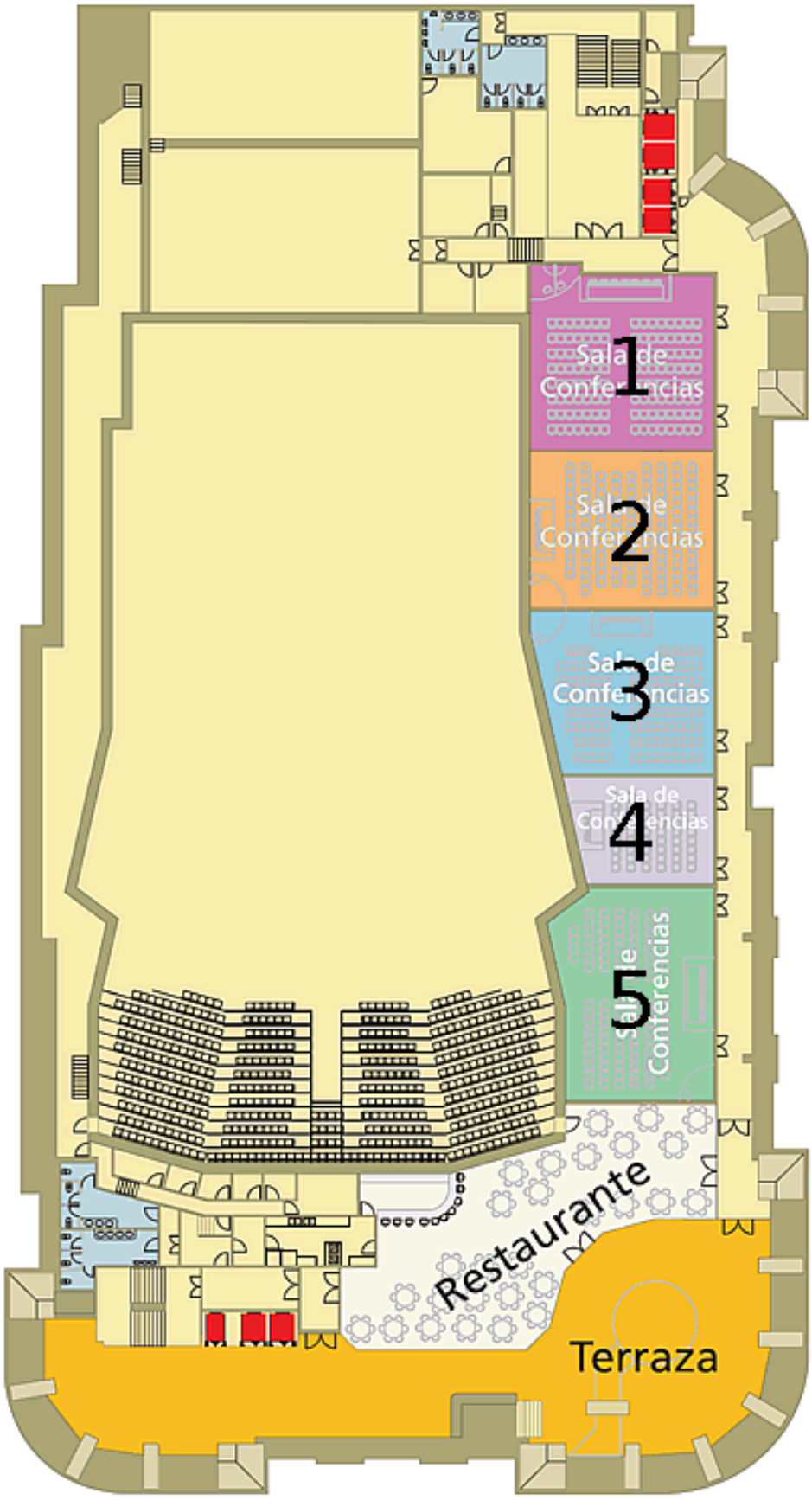


Auditorium/Congress Palace Principe Felipe - Ground Level



Auditorium/Congress Palace Principe Felipe - First Level

Auditorium/Congress Palace Principe Felipe - Third Level



SPONSORS

http://www.cronosaction.com

http://www.iasc-isi.org

http://www.cost.eu

http://www.oviedo.es

http://www.elsevier.com

http://www.uniovi.es

http://www.alquisa.es

http://CMStatistics.org

Contents

Tuesday 23.08.2016 09:40 - 10:30 Room: Sala Camara Chair: Erricos Kontoghiorghes Keynote 1

Trees, forests, and networks

Speaker: **Gerard Biau, Universite Pierre et Marie Curie, France**

Decision tree learning is a popular data-modeling technique used for over fifty years in the fields of statistics, artificial intelligence, and machine learning. The history of trees goes on today with random forests, which are amongst the most successful machine learning algorithms currently available to handle large-scale and high-dimensional data sets. It is sometimes alluded to that forests have the flavor of deep network architectures, insofar as ensemble of trees allow to discriminate between a very large number of regions. However, the connection between random forests and neural networks is largely unexamined. Recent theoretical and methodological developments for random forests will be reviewed, with a special emphasis on the mathematical forces driving the algorithm. Next, the random forest method will be reformulated into a neural network setting, and two new hybrid procedures will be proposed. In a nutshell, given an ensemble of random trees, it is possible to restructure them as a collection of (random) multilayered neural networks, which have sparse connections and less restrictions on the geometry of the decision boundaries. Their activation functions are soft nonlinear and differentiable, thus trainable with a gradient-based optimization algorithm and expected to exhibit better generalization performance.

Thursday 25.08.2016 12:15 - 13:05 Room: Sala Camara Chair: Malgorzata Bogdan Keynote 2

Measuring parameter effects in Bayesian inference

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

A key objective of much of statistical theory concerns the elimination of the effects of nuisance parameters on an inference about an interest parameter. Especially important for statistical practice is quantification of the consequences of including potentially high-dimensional nuisance parameters to provide realistic modelling of a system under study. We consider easily computed measures of nuisance parameter effects in a Bayesian framework. Through decomposition of the Bayesian version of an adjusted likelihood ratio statistic, we propose a computational machinery for analysis of the effects of prior assumptions on nuisance parameters on marginal inference on an interest parameter. Extensions of the techniques allow a formal approach to general sensitivity analysis and evaluations of robustness.

Friday 26.08.2016 12:15 - 13:05 Room: Sala Camara Chair: Gil Gonzalez-Rodriguez Keynote 3

Random objects: Functional data in nonlinear subspaces and Frechet regression

Speaker: **Hans-Georg Mueller, University of California Davis, United States**

Random objects will be illustrated for three commonly encountered scenarios. A general characteristic of random objects is that one has an i.i.d. sample of these objects which lie in a metric space that often has additional properties. The goal is to quantify mean and variation in a sensible way. In the first scenario, functional data that lie on a smooth isometric manifold will be considered. This includes time-warped functional data, where manifold learning with Isomap is shown to provide interpretable data analysis. The second scenario concerns functional data that are density functions. A transformation to a Hilbert space, centered around the Wasserstein mean, then leads to sensible modes of variation. In a third scenario we consider random objects that belong to a more general metric space. We view these as responses in a regression model that features scalar or vector predictors.

Tuesday 23.08.2016 11:00 - 12:50 Parallel Session B – COMPSTAT

MODERN STATISTICAL METHODS FOR COMPLEX DATA

CO087 **Room Sala 3** **Chair: Xinyuan Song**

CO0218: **Analysis of a fixed center effect additive rates model for recurrent event data**

Presenter: **Haijin He**, Shenzhen University, China

Co-authors: Deng Pan, Liang Zhu, Liuquan Sun, Xinyuan Song

A center effect additive rates model is suggested to analyze clustered recurrent event data. The proposed model is a useful alternative to the center effect proportional rates model and provides a direct interpretation of parameters. The traditional estimation methods treat the centers as categorical variables, and they comprise many parameters when the number of centers is large and thus may not be feasible in many situations. An estimation method based on the difference in the observed to the expected number of recurrent events is recommended to address the deficiency of the traditional method. The asymptotic properties of the proposed estimator are established. Simulations are conducted to evaluate the small sample performance and show the computational advantage of the suggested method. The proposed methodology is applied to the Childhood Cancer Survivor study.

CO0245: **Analysis of proportional mean residual life model with latent variables**

Presenter: **Xinyuan Song**, Chinese University of Hong Kong, China

End-stage renal disease (ESRD) is one of the most serious diabetes complications. Numerous studies have been devoted to revealing the risk factors of the onset time of ESRD. We propose a proportional mean residual life (MRL) model with latent variables to assess the effects of observed and latent risk factors on MRL function of ESRD in a cohort of Chinese type 2 diabetic patients. The proposed model generalizes conventional proportional MRL model to accommodate latent risk factors and right censored data. We employ a factor analysis model to characterize latent risk factors via multiple observed variables. We develop a borrow-strength estimation procedure, which incorporates the expectation-maximization algorithm and the corrected estimating equation approach. The asymptotic properties of the proposed estimators are established. Simulation shows that the performance of the proposed methodology is satisfactory. The application to the study of type 2 diabetes reveals insights into the prevention of ESRD.

CO0243: **Bayesian semi-parametric mixed hidden Markov models**

Presenter: **Kai Kang**, The Chinese University of Hong Kong, China

A semi-parametric mixed hidden Markov model is developed to analyze longitudinal data. The proposed model comprises a parametric transition model for examining how potential predictors influence the probability of transition from one state to another and a non-parametric conditional model for revealing the functional effects of explanatory variables on outcomes of interest. Unlike conventional regression that focuses only on the observation process, the proposed model simultaneously investigates the observation process and the underlying transition process. Two correlated random effects, the one is in the conditional model and the other is in the transition model, are considered to describe the possible dependency within and/or between the two stochastic processes. We propose a Bayesian approach that combines Bayesian P-splines and MCMC methods to conduct the statistical analysis. The empirical performance of the proposed methodology is evaluated via simulation studies. An application to a real-life example is presented.

CO0244: **Bayesian local influence of transformation latent variable models with multivariate censored data**

Presenter: **Ming Ouyang**, The Chinese University of HongKong, China

A Bayesian local influence method is developed for transformation latent variable models with multivariate censored data. The effects of minor perturbations to individual observations, the prior distributions of parameters, and the sampling distribution on the statistical inference are assessed through various perturbation schemes. The first-order influence measure is adopted to quantify the degree of minor perturbations to different aspects of a statistical model with the use of Bayes factor as an objective function. Simulation studies show that the empirical performance of the Bayesian local influence procedure is satisfactory.

CO0225: **A relative error-based approach for variable selection**

Presenter: **Meiling Hao**, The Hong Kong Polytechnic University, China

Co-authors: Xingqiu Zhao, Yuanyuan Lin

The accelerated failure time model or the multiplicative regression model is well-suited to analyze data with positive responses. For the multiplicative regression model, an adaptive variable selection method is investigated via a relative error-based criterion and Lasso-type penalty with desired theoretical properties and computational convenience. With fixed or diverging number of variables in regression model,the resultant estimator achieves the oracle property. An alternating direction method of multipliers algorithm is proposed for computing the regularization paths effectively. A data-driven procedure based on the Bayesian information criterion is used to choose the tuning parameter. The finite-sample performance of the proposed method is examined via simulation studies. An application is illustrated with an analysis of one period of stock returns in Hong Kong Stock Exchange.

ARS-IASC SESSION I: NEW COMPUTATIONAL APPROACHES TO NONLINEARITY, DIMENSION REDUCTION AND CLUSTERING

CO031 **Room Sala 1** **Chair: Yuichi Mori**

CO0231: **Predictive clustering using a component-based approach**

Presenter: **Michio Yamamoto**, Kyoto University, Japan

Co-authors: Atsushi Kawaguchi, Heungsun Hwang

A novel clustering method is proposed to identify a cluster structure that is related to outcome variables and to predict cluster memberships of future individuals based on a large number of explanatory variables. Technically, the proposed method carries out outcome-guided dimension reduction (ODR) and the clustering of the dimension-reduced subspace simultaneously. ODR is proven to coincide with a type of partial least squares (PLS) regression, indicating that the proposed method represents a simultaneous approach to PLS and subspace-clustering. PLS is a component-based approach, in which components are defined as linear combinations of explanatory variables. Thus, the combined method can help a researcher to interpret which explanatory variables have effects on the cluster structure based on the weight in linear combinations. In addition, sparse estimation of weights is adopted to obtain a more interpretable and stable cluster structure. Simulated and real data analyses show that the proposed method can provide a cluster structure that is associated with outcome variables and predict cluster memberships of future individuals well.

CO0234: **Integrating multiple random sketches of singular value decomposition**

Presenter: **Su-Yun Huang**, Academia Sinica, Taiwan

Low-rank singular value decomposition (SVD) of large-scale matrices is a key tool in modern data analysis and scientific computing. Rapid growing in matrix size further increases the needs and poses the challenges for developing efficient large-scale SVD algorithms. Random sketching is a promising method to reduce the problem size for computing an approximate SVD. We generalize the one-time sketching to multiple random sketches and develop algorithms to integrate these random sketches containing various subspace information in different randomizations. Such integration procedure can lead to SVD with higher accuracy and the multiple randomizations can be conducted on parallel computers simultaneously.

We also reveal the insights and analyze the performance of the proposed algorithms from statistical and geometric viewpoints. Numerical results are presented and discussed to demonstrate the efficiency of the proposed algorithms.

CO0261: **Canonical covariance analysis for three-mode three-way data by using connector matrix**
Presenter: **Jun Tsuchida**, Doshisha University, Japan
Co-authors: Hiroshi Yadohisa

Given two three-mode three-way data sets, such as panel data sets, two types of factors are generally investigated: common factors, which show relationships between the two data sets, and unique factors, which represent the uniqueness of each data set. We propose a method for investigating the common and unique factors simultaneously. Canonical covariance analysis is one such method; however, this method has been proposed for two-mode two-way data and regards the same variable under different conditions as being two different variables. Moreover, this method makes it difficult to distinguish between the two types of factors because parameters of common factors are not separated from parameters of unique factors. To address these problems, we introduce a connector matrix and extend the canonical covariance analysis from two-mode two-way to three-mode three-way data sets. Using a connector matrix makes it easy to distinguish between the two types of factors. Furthermore, we can choose different numbers of dimensions for the canonical variable between two data sets.

CO0304: **Acceleration of iterative methods for nonnegative matrix factorization**
Presenter: **Yuichi Mori**, Okayama University of Science, Japan
Co-authors: Michio Sakakihara, Masahiro Kuroda, Masaya Iizuka

Nonnegative matrix factorization (NMF) is applied to several problems in data analysis, for example, clustering, pattern recognition and multimedia data analysis. The alternating least squares (ALS) algorithm is a simple iterative method to give us a factorization in the sense of Frobenius norm of matrices. When applying the ALS algorithm to NMF of large scale data matrix, the algorithm converges slowly because its convergence is almost linear. We propose two component-wise acceleration methods of the ALS algorithm for improving its rate of convergence. These acceleration methods are based on a two point acceleration scheme called Aitken delta-squared method and a three point acceleration scheme by the use of a rational interpolation. The three point acceleration scheme has a better convergence property than the two point acceleration scheme. We prove the fast convergence property of the proposed accelerations under the convergence condition of the ALS algorithm, and evaluate the effectiveness of the proposed accelerations for NMF in numerical experiments.

CO0227: **Some mathematical notes on comprehensive factor analysis**
Presenter: **Kohei Adachi**, Osaka University, Japan
Co-authors: Nickolay Trendafilov

In the currently prevalent model of factor analysis (FA), specific factors and errors are not dissociated, though they had been separated in the original conception of FA. Thus, an FA model with specific factors dissociated from errors is considered, whose least squares procedure is referred to as comprehensive FA. The main goal includes showing that the model part of comprehensive FA and residuals are identifiable, common and specific factors are undetermined but have some constancy, and comprehensive FA has clear relationships with principal component analysis.

TUTORIAL 1

CO100 **Room Sala Camara** **Chair: Christine Keribin**

CO0578: **Cluster validation: How to think and what to do**
Presenter: **Christian Hennig**, UCL, United Kingdom

Cluster analysis is about finding groups in data. There are many cluster analysis methods and on most datasets clusterings from different methods will not agree. Cluster validation concerns the evaluation of the quality of a clustering. This is often used for comparing different clusterings on a dataset, stemming from different methods or with different parameters such as the number of clusters. An overview will be given of techniques for cluster validation, including visualisation methods, methods for assessing stability if a clustering, tests, validity indexes and some new measurements of different aspects of cluster validity. A discussion will be made on the issue of what the "true clusters" are that we want to find and how this depends on the specific application and the aims and concepts of the researcher, so that these can be connected to specific techniques for cluster validation. In the literature, the problem of cluster validation is often not well defined and there is a focus on automatic methods without providing much understanding of the specific circumstances in which they work (or not). Some insight into these issues will be provided.

ROBUST STATISTICS I

CG022 **Room Sala 2** **Chair: Agustin Mayo-Iscar**

CC0156: **Efficient computation of the minimum weighted covariance determinant estimator**
Presenter: **Jan Kalina**, Faculty of Social Sciences Charles University in Prague, Czech Republic
Co-authors: Jurjen Duintjer Tebbens

We study efficient algorithms for robust estimators of multivariate location and scatter. First, we propose an efficient algorithm for the existing Minimum Weighted Covariance Determinant (MWCD) estimator, which can be interpreted as a weighted analogy of the popular Mininum Covariance Determinant (MCD) estimator. The algorithm exploits suitable tools of numerical linear algebra. Further, we propose a new version of the estimator, which is again based on implicit weights assigned to individual observations, but an original idea allows the estimator to be computationally much more efficient. Our theoretical results include the asymptotic efficiency and breakdown point derived for the novel method. Nevertheless, the main contribution is again a proposal of an efficient algorithm for its computation.

CC0364: **Robust estimation and moment selection in dynamic fixed-effects panel data models**
Presenter: **Pavel Cizek**, Tilburg University, Netherlands
Co-authors: Michelle Aquaro

The aim is to extend an existing outlier-robust estimator of linear dynamic panel data models with fixed effects, which is based on the median ratio of two consecutive pairs of the first-differenced data. To improve its precision and robust properties, a general procedure based on many pairwise differences and their ratios is designed. The evaluated ratios are combined by means of the proposed two-step GMM estimator, which newly relies on the weighting scheme reflecting both the variance and bias of the moment equations; the bias is assumed to stem from the outliers and data contamination and needs to be numerically approximated and estimated. Additionally, a robust criterion for selecting the number of moment conditions is proposed. The asymptotic distribution as well as robust properties of the estimator are derived; the latter are obtained both under contamination by independent additive outliers and the patches of additive outliers. The proposed estimator is additionally compared with existing methods by means of Monte Carlo simulations.

CC0389: **Power properties of the out-of-sample portfolio performance measures tests**
Presenter: **Ekaterina Kazak**, University of Konstanz, Germany
Co-authors: Winfried Pohlmeier

While the literature on portfolio choice largely concentrated on stabilization strategies, little attention has been devoted to the quality of performance tests used to check, if a given strategy can significantly outperform an alternative in terms of some performance measure. We examine the quality of portfolio performance measures tests and conclude that the puzzling empirical results of inferior performance of the theoretically superior

strategies based on the out-of-sample comparison are coming partly from the low power properties of the tests. We emphasize the importance of the underlying return distribution and show that the out-of-sample portfolio returns follow a mixture distribution depending on the return vector, but also the estimated portfolio weights. In the simulation study with the proposed mixture distribution design we show that in the realistic cases the test difference is overemphasized, the main issue here is the low testing power, which automatically leads to a conclusion, that the benchmark strategy cannot be outperformed significantly. More specifically, we show that in some circumstances it might be reasonable to select a low significance level (high Type One error) and to choose the alternative rather than sticking to the model that cannot be rejected under the null.

CC0430: **Nonparametric permutation-based testing and ranking on multivariate time data**
Presenter: **Livio Corain**, University of Padova, Italy

Based on the concept of multivariate stochastic dominance, the aim is to propose a nonparametric and permutation-based method for testing and ranking on multivariate time data. By using either fixed or moving blocks, the proposed methodology provide a flexible and less demanding in terms of underlying assumptions tool to infer on the presence of possible stochastic dominances that may take place among a set of several multivariate populations. Via a Monte Carlo simulation study, we investigate the properties of the proposed testing and ranking method where we prove its validity under different random distributions and type of dependencies and correlation structures. From the practical point of view, the proposed methodology can be effective to face some real problems in Econometrics and Finance. Finally, we present and application to a macroeconomic analysis of an industrial sector.

CC0237: **Shrinkage estimation through the ridGME procedure**
Presenter: **Pedro Macedo**, University of Aveiro, Portugal

The importance of ridge regression in the estimation of ill-conditioned models is well-known. Recently, the ridGME procedure appears in the literature as a simple and straightforward shrinkage estimation technique with a good performance. In addition to theoretical background, the algorithm and its implementation in MATLAB software will be discussed in order to illustrate the ridge regression procedure.

COPULAS

CG052 **Room Sala 5** **Chair: Anouar El Ghouch**

CC0355: **Estimating copula density: Convergence speed results**
Presenter: **Jerome Collet**, Electricite de France RD Division, France

The goal is to estimate the copula density of a $d-$dimensional random variable, without parametric assumptions, using ranks and subsampling. The main feature of this method is a low sensitivity to dimension, on realistic cases. We give a description of the estimation method, a convergence proof, and some hints on convergence speed. In the simple useless case of independence, we prove the convergence speed is the same as for kernel density estimation. In more structured cases, numerical simulations show the impact of dimension is much smaller than for a kernel estimation. Furthermore, we prove that if the underlying model is mainly additive, the impact of dimension is the same as in the parametric case.

CC0458: **Semiparametric copula quantile regression for complete or censored data**
Presenter: **Anouar El Ghouch**, The University catholique de Louvain, Belgium

When facing multivariate covariates, general semiparametric regression techniques come at hand to propose flexible models that are unexposed to the curse of dimensionality. A semiparametric copula-based estimator for conditional quantiles is investigated for complete or right-censored data. Extending recent work, the main idea consists in appropriately defining the quantile regression in terms of a multivariate copula and marginal distributions. Prior estimation of the latter and simple plug-in lead to an easily implementable estimator expressed, for both contexts with or without censoring, as a weighted quantile of the observed response variable. In addition, and contrary to the initial suggestion in the literature, a semiparametric estimation scheme for the multivariate copula density is studied, motivated by the possible shortcomings of a purely parametric approach and driven by the regression context. The resulting quantile regression estimator has the valuable property of being automatically monotonic across quantile levels, and asymptotic normality for both complete and censored data is obtained under classical regularity conditions. Finally, numerical examples as well as a real data application are used to illustrate the validity and finite sample performance of the proposed procedure.

CC0493: **How to estimate CoVaR**
Presenter: **Piotr Jaworski**, University of Warsaw, Poland

CoVaR (conditional Value at Risk) is a newly introduced risk measure which is oriented on systemic risk. If random variables X and Y are modelling our phenomena, say welfares of banks or gains from the investments, CoVaR of Y with respect to X is VaR of conditional Y, conditioned on the poor standing of X. In more details, $CoVaR_\beta(Y|X) = VaR_\beta(Y|X \in E)$, where E, the Borel subset of the real line, is modelling some adverse event concerning X. The basic properties of CoVaR, especially those which depend on the copula of the pair X,Y, and some methods of its estimation on the base of empirical data will be presented.

CC0450: **Model distances for vine copulas in high dimensions**
Presenter: **Matthias Killiches**, Technische Universitaet Muenchen, Germany
Co-authors: Daniel Kraus, Claudia Czado

Vine copulas are a flexible class of dependence models consisting of bivariate building blocks and have proven to be particularly useful in high dimensions. Classical model distance measures require multivariate integration and thus suffer from the curse of dimensionality. We provide numerically tractable methods to measure the distance between two vine copulas even in high dimensions. For this purpose, we consecutively develop three new distance measures based on the Kullback-Leibler distance, using the result that it can be expressed as the sum over expectations of KL distances between univariate conditional densities, which can be easily obtained for vine copulas. To reduce numerical calculations we approximate these expectations on adequately designed grids, outperforming Monte Carlo-integration with respect to computational time. In numerous examples and applications we illustrate the strengths and weaknesses of the developed distance measures.

CC0437: **Representing sparse Gaussian DAGs as sparse R-vines allowing for non-Gaussian dependence**
Presenter: **Dominik Mueller**, Technische Universitaet Muenchen, Germany
Co-authors: Claudia Czado

Modeling dependence in high dimensional systems has become an increasingly important topic. Most approaches rely on the assumption of a joint Gaussian distribution such as statistical models on directed acyclic graphs (DAGs). They are based on modeling conditional independencies and are scalable to high dimensions. In contrast, vine copula based models can accommodate more elaborate features like tail dependence and asymmetry. This flexibility comes however at the cost of exponentially increasing complexity for model selection and estimation. We show a connection between these two model classes by giving a novel representation of DAG models in terms of sparse vine copula models. Therefore, we can exploit the fast model selection and estimation of sparse DAGs while allowing for non-Gaussian dependence in the vine copula models. We evaluate our methodology by a large scale simulation studiy and high dimensional data examples demonstrating that our approach outperforms standard methods for vine copula structure selection.

COMPUTATIONAL STATISTICS FOR CATEGORICAL DATA

CG080 **Room Sala 7** **Chair: Luca La Rocca**

CC0518: **Inverse multiple correspondence analysis**

Presenter: **Michel van de Velden**, Erasmus University Rotterdam, Netherlands

Co-authors: Wilco van den Heuvel, Patrick Groenen

In correspondence analysis (CA), the aim is to obtain a low-dimensional representation that optimally depicts associations in a two-way contingency table. An important and useful extension of CA is multiple correspondence analysis (MCA). MCA allows a simultaneous representation of the observations (subjects) and the categories of several categorical variables in a space of, user selected, low-dimensionality. For inverse MCA, this low-dimensional solution is given and the aim is to retrieve the original data or, if such is not possible, to identify possibly underlying data sets. In previous work on inverse methods for CA, it was shown that solutions can be found by taking convex combinations of a set of so-called vertices. Moreover, for problems where the original data matrix is of relatively small dimensionality, the complete set of such vertices can be obtained using complete enumeration procedures. However, the proposed methods cannot be applied to larger problems, and are therefore not suited for inverse MCA problems. We take a different approach to the inverse CA and MCA problems, by considering it from an integer programming perspective.

CC0506: **An optimal nearest neighbor hot deck imputation based on a b-matching problem**

Presenter: **Dennis Kreber**, Trier University, Germany

Co-authors: Jan Pablo Burgard, Sven de Vries, Ulf Friedrich

Almost all population surveys suffer from missing responses, inhibiting the direct application of estimation methods requiring complete data sets. In official statistics usually single imputation methods are applied to create one complete and coherent data set. A prominent single imputation variant is the nearest neighbor hot deck imputation. It replaces the missing values with observed values from the closest donor given a distance, e.g. the Gower distance. A repeated assignment of donors to non-respondents may lead to distortions in the distribution of the data. To avoid such a problem the proposed method allows to limit the number of maximum donations per unit. Then the sum over all distances between imputation pairs is minimized globally. This leads to a maximum weighted b-matching problem that is solved exactly by a combinatorial optimization procedure. The proposed method is compared to existing single imputation methods within a large scale Monte-Carlo simulation based on the Amelia dataset. The estimation of a total is studied under different missing patterns for a variety of variables, e.g. income. Variance estimation is performed via bootstrap. The Monte Carlo simulation indicates that the imposed restriction on the reuse of donors may lead to a lower bias and variance.

CC0544: **An unconstrained approach to rotational indeterminacy in Bayesian exploratory multidimensional IRT models**

Presenter: **Sara Fontanella**, The Open University, United Kingdom

Co-authors: Lara Fontanella, Pasquale Valentini, Nickolay Trendafilov

Within the social and behavioural sciences, item-level data are often categorical in nature and item factor analysis (IFA) represents an appropriate tool for their analysis. We consider only a specific class of factor analytic models, namely Multidimensional Item Response Theory (MIRT) models. These models can be defined in terms of both exploratory and confirmatory perspectives. In the former context, identification problems have to be considered. We focus on the rotational indeterminacy. Following a Bayesian perspective, we address this issue by considering an ex-ante approach, which imposes a minimal number of constraints on the model parameters, as well as an ex-post approach, proposed in classical exploratory factor analysis. Specifically, the first method represents a constrained version of MIRT models where the rotation indeterminacy is removed by imposing a positive lower triangular structure on the factor loadings matrix. On the contrary, the ex-post procedure relies on the definition of an unconstrained Gibbs sampler where the rotational invariance is addressed in a post-processing procedure based on the Orthogonal Procrustes approach. However, in the context of MIRT models, one has to take into account the correlations between the latent traits. For this reason, the post-processing procedure is replaced by Oblique Procrustes.

CC0446: **Visualizations of multiple imputations using generalized orthogonal procrustes analysis**

Presenter: **Johane Nienkemper-Swanepoel**, Stellenbosch University, South Africa

Co-authors: Niel Le Roux, Sugnet Lubbe

Multiple imputation based on multiple correspondence analysis (MIMCA) has been suggested for dealing with missing values in categorical data sets. The MIMCA procedure is visually investigated using simulated data sets with different patterns of missing data. Multiple correspondence analysis (MCA) biplots of the multiple imputations are constructed and optimally aligned using Generalized orthogonal Procrustes analysis (GOPA). GOPA allows the comparison of several configurations with a group average configuration or predetermined target configuration. The aligned biplots allow a detailed description of the consistencies and idiosyncrasies among the various imputed data sets. An average configuration can be obtained from the optimally aligned configurations, which is intuitively associated to the well-known combination rules of Rubin used for combining estimates from multiple imputed data sets. It is proposed to use the distances between the samples and category level points (CLPs) of the average configuration to predict final CLPs for the missing samples. This proposal will result in a final combined data set for further analysis, as opposed to multiple data sets with separate analyses. Finally visualizations of the predicted data set are compared to visualizations of the original complete simulated data set. Different measures for the goodness of fit within the Procrustes framework will be used to validate the proposed procedures.

CC0372: **A new composite indicator for sensorial data**

Presenter: **Stefano Bonnini**, University of Ferrara, Italy

The studies on the environmental impact of bad odors are based on different approaches and methods: atmospheric dispersion models, olfactometry, chemical assessments by means of electronic noses and sniffing team surveys. The studies based on sniffing teams are preferable when the focus is on the perceived odors and in the presence of several possible sources of bad odors, because with the chemical methods it is difficult to distinguish different types of odors and to detect odors with low olfactive threshold. A sniffing team is a group of trained panellists engaged to produce field measurements of odor perceptions. All the mentioned approaches agree with the idea that the impact of perceived odors mainly depends on the duration and on the intensity of the perceptions. A composite indicator to measure malodour annoyance in a sniffing team survey is proposed. The new index jointly consider duration and intensity of the perceptions. Its application on the data of a survey performed in 2010 in an area in the north of Italy, shows its usefulness in the assessment of regional differences and trend over the time of odor perceptions.

PATTERN RECOGNITION OF TIME SERIES

CG028 **Room Sala 4** **Chair: Peter Tillmann**

CC0548: **Combining structural change analysis with anomaly detection: A constrained clustering approach**

Presenter: **Carlo Drago**, University of Rome Niccolo Cusano, Italy

The identification of the structural changes is a very important problem in modern time series analysis. We integrate the structural change analysis with the anomaly detection also defined as the specific identification of the deviations from a given pattern. It is proposed an approach for decomposing the time series in components and identifying the different structural breaks of the time series. In this sense it is considered a constrained hierarchical clustering algorithm in order to detect the anomalies and the structural changes over time. The approach is relevant in order to simultaneously determine the structure of the time series, the structural breaks and the different anomalies (single or subsequent observations) over time. The anomalies can occur considering the direction and the relevance of the structural changes over time. The approach is investigated by using simulated and also real time series.

CC0550: **Capturing correlation changes by applying kernel change point detection on the running correlations**

Presenter: **Jedelyn Cabrieto**, KU Leuven, Belgium

Co-authors: Francis Tuerlinckx, Peter Kuppens, Eva Ceulemans

Change point detection methods signal the occurrence of abrupt changes in a time series. Non-parametric approaches are especially attractive in this regard because they impose less assumptions on the data. Yet, a drawback of these methods is that they are expected to be sensitive to changes in the mean, the variance, the correlation, and even the higher moments. This implies that one is not certain what kind of change the methods pick up, whereas this is often important from a substantive point of view. We focus on signaling correlation change, because this is put forward by different theories but proved hard to trace in multivariate time series. We demonstrate how correlation change can be detected by applying the best non-parametric method, kernel change point detection, on the running correlations rather than on the raw data. We inspect the detection performance of this approach in a simulation study and provide an illustrative example on detecting synchronicity in reactivity data.

CC0383: **Time series forecasting with a learning algorithm: An approximate dynamic programming approach**

Presenter: **German Creamer**, Stevens Institute of Technology, United States

Co-authors: Ricardo Collado

The focus is on the basic problem of re-fitting a time series over a finite period of time and formulate it as a stochastic dynamic program. By changing the underlying Markov decision process we are able to obtain a model that at optimality considers historical data as well as forecasts of future outcomes. We design lookahead dynamic methods for the solution of our Markov decision process. By recursively applying this idea over a range of future time periods, look-ahead dynamic programming methods effectively react to changes in the data and consider the stream of future outcomes obtained from our model decisions. Employing these techniques should give models calibrated to historical data which at any point in time would be optimally positioned to react to possible future data stream.

CC0563: **A deep learning approach for electrical signal time series classification**

Presenter: **Cem Iyigun**, Middle East Technical University, Turkey

Various methods have been studied in time series classification over past decades and successful results have been achieved in many aspects. But, there is still room for improvement. Noisy, high dimensional and complex time series data cannot be modeled with traditional shallow methods that have limited non-linear operation ability. We attempt to classify electrical signal based time series data by using deep learning algorithms. Deep learning based proposed method will be compared with one commonly used recurrent neural network algorithm which is specialized for time series modeling, multi class SVM based methods with various kernels and nearest neighbor algorithm with dynamic time warping based distance measures.

CC0406: **Forecasting financial time big data using interval time series**

Presenter: **Carlos Mate**, Universidad Pontificia Comillas, Spain

Co-authors: Javier Redondo

An interval time series (ITS) assigns to each period an interval covering the values taken by the variable. Each interval has four characteristic attributes, since it can be defined in terms of lower and upper boundaries, center and radius. The analysis and forecasting of ITS is a very young research area, dating back less than 15 years, and still presents a wide array of open issues. One main issue with time series in a big data context consists of deciding if to handle it as classic time series (CTS) or to proceed with some kind of aggregation in order to get a time series of symbolic data like ITS. Using the k-Nearest Neighbours (kNN) method, both approaches are applied to forecast exchange rates. Based on usual distances for interval-valued data such as Haussdorff, Ichino-Yaguchi and so on; the reduction in mean distance error using ITS instead of CTS suggests that the ITS approach could be a better way to forecast exchange rates using large data or data streaming. Some interesting conclusions about monthly and daily aggregation horizons are obtained and further research issues are proposed.

Tuesday 23.08.2016 14:30 - 16:00 Parallel Session C – COMPSTAT

BAYESIAN NONPARAMETRICS

CI073 **Room Sala Camara** **Chair: Simone Padoan**

CI0182: **Semiparametric Bayesian regression via Potts model**
Presenter: **Fernando Quintana**, Pontificia Universidad Catolica de Chile, Chile
Co-authors: Alejandro Murua

Bayesian nonparametric regression through random partition models is considered. Our approach involves the construction of a covariate-dependent prior distribution on partitions of individuals. Our goal is to use covariate information to improve predictive inference. To do so we propose a prior on partitions based on the Potts clustering model associated with the observed covariates. This drives by covariate proximity both the formation of clusters, and the prior predictive distribution. The resulting prior model is flexible enough to support many different types of likelihood models. We focus the discussion on nonparametric regression. Implementation details are discussed for the specific case of multivariate multiple linear regression. The proposed model performs well in terms of model fitting and prediction when compared to other alternative nonparametric regression approaches. We illustrate the methodology with an application to the health status of nations at the turn of the 21st century.

CI0250: **Hierarchical processes for Bayesian nonparametric inference**
Presenter: **Antonio Lijoi**, University of Pavia and Collegio Carlo Alberto, Italy
Co-authors: Federico Camerlenghi, Peter Orbanz, Igor Pruenster

Recent findings are discussed on the use of completely random measures for constructing priors suitable for Bayesian nonparametric inference with data that display dependence structures more general than exchangeability. The focus will be on a broad class of hierarchical processes whose distributional properties will be presented. These theoretical results are the key ingredients for devising suitable MCMC algorithms that may rely either on a "marginal" or on a "conditional" approach. Illustrations related to prediction within species sampling problems and inference on survival data will be finally displayed.

CI0256: **Sparse and modular networks using exchangeable random measures**
Presenter: **Francois Caron**, University of Oxford, United Kingdom

Statistical network modeling has focused on representing the graph as a discrete structure, namely the adjacency matrix, and considering the exchangeability of this array. In such cases, it is well known that the graph is necessarily either dense (the number of edges scales quadratically with the number of nodes) or trivially empty. We instead consider representing the graph as a measure on the plane. For the associated definition of exchangeability, we rely on the Kallenberg representation theorem. For certain choices of such exchangeable random measures underlying the graph construction, the network process is sparse with power-law degree distribution, and can accommodate an overlapping block-structure. We then present a Markov chain Monte Carlo algorithm for efficient exploration of the posterior distribution and demonstrate that we are able to recover the structure of a range of networks graphs ranging from dense to sparse based on our flexible formulation.

ARS-IASC SESSION II: COMPUTATIONAL ALGORITHMS TO JOINT INFERENCE ON DESIGN AND MODELING

CO043 **Room Sala 1** **Chair: Yi-Ting Hwang**

CO0179: **Smart watch unboxing video viewing frequency prediction by support vector machine and Bayesian probit analysis**
Presenter: **Nai-Hua Chen**, Chienkuo Technology University, Taiwan

The quick changing technology produces make innovation products filled in the marketplace to compete customers intention. The unboxing articles reveal through blogs describing details of new purchased products. Due to the development of media techniques, unboxing process are shared in video type via Youtube. It provides viewers an unboxing-like experience while watching and becomes more popular. These videos help industries to explosure their newly launched products. Some researchers have devoted in modeling the consumer innovation adoption behavior. Viewers behavior towards unboxing videos is associated with their preferences. It can adopt to understand whether customers attitude towards the new products. Several purchase models are developed to understand customer dynamic behavior. The support vector machine (SVM) has gained popularity in visual pattern recognition recently. The negative binomial distribution (NBD) contains two characteristics which are Poisson purchasing and gamma heterogeneity is shown as a robust method in product purchase frequency. We proposed a two-stage innovation adoption model. The first is to classify unboxing video viewers into like and unlike. Next, the Bayesian probit analysis is used to compare effects that influence like and unlike behavior.

CO0195: **Evaluating human networks of author's affiliation info and co-author info in scientific literature by using centralities**
Presenter: **Yuji Mizukami**, Nihon University, Japan
Co-authors: Yosuke Mizutani, Keisuke Honda, Shigenori Suzuki, Junji Nakano

The Institute of Statistical Mathematics aims to promote the joint use of resources and assets by domestic and foreign researchers, so as to further develop the statistics. We discuss it from a statistical point of view. A new index is required, by which the progress and effectiveness of the joint use can be objectively assessed. We analyze the co-authored information in extended betweenness centrality in papers that were written by researchers of the Institute of Statistical Mathematics in Web of Science. The results of the analysis make it possible to identify the researched people that are in the center of each research field. Our analytical activity consists of two stages. In the first analytical stage (stage 1), we use betweenness centrality to identify people that are in the center of each research field (network sector). In the next analytical stage (stage 2), we proposed four new analysis formulas that were based on betweenness centrality to analyze separately the researcher networks inside the organization and networks within and outside the organizations.

CO0232: **The joint model for the survival data and binary repeated measures**
Presenter: **Yi-Ting Hwang**, National Taipei University, Taiwan
Co-authors: Chia-Hui Huang, Chun-Chao Wang, Yi-Kang Tseng

The medical cost in an aging society will increase substantially if the elderly have higher incidence of chronic diseases, disability and unable to live independently. Healthy lifestyle not only affects elderly individuals but also influence the entire community. When assessing the healthy lifestyle, survival and quality of life should be considered concurrently. Thus, simultaneously identifying the association between the survival and long-term quality of life becomes an important issue. Jointly modeling two outcomes have been studied previously. Most of the existing models have a sequence of continuous repeated measurements, which are modeled by the general linear model for longitudinal data. A modified joint model for modeling survival and the longitudinal binary repeated measures simultaneously is proposed. The joint likelihood estimation is used. Owing to some unobservable information in the model, some parameters in joint model have to be estimated by Monte Carlo EM algorithm and Metropolis-Hastings algorithm. Monte Carlo simulations are used to evaluate the performance of the proposed model based on the accuracy and precision of the estimates. A real data is used to illustrate the feasibility of the proposed model.

CO0246: **Economic design of two-stage control charts with asymmetric and dependent measurements**
Presenter: **Nan-Cheng Su**, National Taipei University, Taiwan

In many instances, the cost is high to monitor primary quality characteristic called performance variable, but it could be more economical to monitor

its surrogate. To cover asymmetric processes for two-stage charting methods using both performance and surrogate variables, bivariate skew normal distribution is considered as the underlying distribution of process variables. When the correlation relationship between the performance variable and its surrogate is specified, a charting procedure to monitor either the performance variable or its surrogate in an alternating fashion rather than monitoring the performance variable alone is proposed. The proposed two-stage control charts are constructed under an economic design using Markov chain approach. Two algorithms are provided to implement the proposed charting method. The application of the proposed charting method and its advantages over the existing methods are presented through an illustrating example.

ROBUSTNESS IN REGULARIZED PROBLEMS

CO025 **Room Sala 2** **Chair: Christophe Croux**

CO0238: **Robust regularised precision matrix estimation**

Presenter: **Garth Tarr**, University of Newcastle, Australia

There is a great need for robust techniques in data mining and machine learning contexts where many standard techniques such as principal component analysis and linear discriminant analysis are inherently susceptible to outliers. Furthermore, standard robust procedures assume that less than half the observation rows of a data matrix are contaminated, which may not be a realistic assumption when the number of observed features is large. We consider the problem of estimating covariance and precision matrices under cellwise contamination. Specifically, the use of a robust pairwise covariance matrix as an input to various regularisation routines, such as the graphical lasso, QUIC or CLIME. We review a number of approaches that can be used to ensure the input covariance matrix is positive semidefinite. The result is a potentially sparse precision matrix that is resilient to moderate levels of cellwise contamination and scales well to higher dimensions. We consider the selection of an appropriate value for the tuning parameter that controls the level of sparsity and potential applications involving financial data and bioinformatics.

CO0290: **Sparse S- and MM-estimation for high-dimensional regression**

Presenter: **Andreas Alfons**, Erasmus University Rotterdam, Netherlands

Co-authors: Christophe Croux, Viktoria Oellerer

The S-estimator and the MM-estimator are among the most popular robust regression estimators. While both estimators are highly robust against outliers in the data, the latter has the advantage of also attaining high efficiency. However, those methods cannot be applied to high-dimensional data, i.e. data with more variables than observations. As a remedy, the sparse S-estimator and the sparse MM-estimator are defined by adding an L1 penalty on the coefficient estimates to the respective objective functions. These new estimators combine robust regression with sparse model estimation, but fast algorithms are required for use in practical applications. In addition to presenting algorithms for the computation sparse S and sparse MM, their performance is assessed by means of a simulation study.

CO0346: **Sparse robust regression estimators**

Presenter: **Gabriela Cohen Freue**, University of British Columbia, Canada

Co-authors: David Kepplinger, Matias Salibian-Barrera

In many current applications scientists can easily measure a very large number of variables (for example, several thousands of gene expression levels) some of which are expected be useful to explain or predict a specific response variable of interest. These potential explanatory variables are most likely to contain redundant or irrelevant information, and in many cases, their quality and reliability may be suspect. We developed a penalized robust regression estimator that can be used to identify a useful subset of explanatory variables to predict the response, while protecting the resulting estimator against possible aberrant observations in the data set. Using an Elastic Net penalty, the proposed estimator can be used to select variables, even in cases with more variables than observations or when many of the candidate explanatory variables are correlated. We present the new estimator and an algorithm to compute it. We also illustrate performance of the proposed estimator in a simulation study and a real data set.

CO0342: **Initial estimators for regularized robust methods in high-dimensional settings**

Presenter: **David Kepplinger**, University of British Columbia, Canada

Co-authors: Matias Salibian-Barrera, Gabriela Cohen Freue

Many robust methods involve the minimization of a non-convex function, thus the initial value for the optimization algorithm is of particular important to attain a good solution. We compare the performance of initial estimators for regularized methods when there are more parameters than observations. Although random subsampling is the prevalent approach to get initial estimates, deterministic initial estimates are increasingly popular alternatives. With regularized methods, the number of subsamples does not have to increase with the dimensionality of the problem, but many subsamples must be considered nevertheless. Other methods generate subsamples in more informed ways to reduce the number of subsamples. We consider Principal Sensitivity Components-generalized to regularized estimators for high-dimensional problems-to guide the search for good subsamples. Since regularized estimators are generally computed over a grid of penalty values, a potential estimator to start the optimization at a given penalty level is the when there are to a previous optimization with a very similar degree of regularization. This includes the particularly useful approach to compute the regularization path starting from an all-zero coefficient and gradually relaxing the penalty, which has the benefit of avoiding a cold start. We compare these different initial estimates for robust regularized S-estimates of regression in terms of the attained objective function, sparsity, and computational speed.

TIME-VARYING COEFFICIENTS

CG013 **Room Sala 3** **Chair: Michael Smith**

CC0167: **Empirical likelihood based inference for fixed effects varying coefficient panel data models**

Presenter: **Luis Antonio Arteaga Molina**, Universidad de Cantabria, Spain

Co-authors: Juan Manuel Rodriguez-Poo

Local empirical likelihood based inference for non parametric varying coefficient panel data models with fixed effects is investigated. First, we show that the naive empirical likelihood ratio is asymptotically standard chi-squared when undersmoothing is employed. The ratio is self-scale invariant and the plug-in estimate of the limiting variance is not needed. Second, mean-corrected and residual-adjusted empirical likelihood ratios are proposed. The main interest of these techniques is that without undersmoothing, both also have standard chi-squared limit distributions. As a by-product, we propose also two empirical maximum likehood estimators of the varying coefficient models and their derivatives. We obtain as well the asymptotic distribution of these estimators. Furthermore, a non parametric version of the Wilk's theorem is derived. To show the feasibility of the technique and to analyze its small sample properties, using empirical likelihood-based inference we test for a conditional factor model in the CAPM setting and we implement a Monte Carlo simulation exercise.

CC0188: **Time-varying SURE: A nonparametric approach**

Presenter: **Isabel Casas**, BCAM, Spain

Co-authors: Eva Ferreira, Susan Orbe

A local linear estimator is proposed for a SUR model (LL-SUR) with time-varying coefficients. The asymptotic results, consistency and asymptotic normality, are obtained under locally stationary variables. Contrary to the parametric case, the results show that the efficiency of LL-SUR outperforms the LL estimator separately estimated for each equation even when the covariates are the same for all equations. A simulation study

is conducted to show the performance of the methodology in finite samples. Moreover, a discussion and practical procedure is derived to select the smoothing parameters. Finally, we present an application to estimate time-varying coefficients in beta pricing models.

CC0516: **Integer-valued autoregressive models with dynamic coefficient driven by a stochastic recurrence equation**
Presenter: **Paolo Gorgi**, University of Padua, Italy

A new class of integer-valued autoregressive (INAR) models with dynamic coefficient is proposed. The peculiarity of this class of models lies in the specification of the INAR coefficient through a stochastic recurrence equation. The estimation of the model can be performed by maximum likelihood and the consistency of the estimator is proved. The flexibility of the proposed specification is illustrated in a simulation study. An application to a time series of crime reports is presented. The results show how the dynamic coefficient can allow to enhance both the in-sample and the out-of-sample performance of INAR models.

CC0461: **A linear mixed-effect state space model to discriminate unobservable structural components**
Presenter: **Marco Costa**, University of Aveiro, Portugal
Co-authors: Magda Monteiro

A linear mixed-effect state space model is proposed in order to discriminate monthly environmental time series through the prediction of unobservable components (for instance, trend and seasonality components). This model incorporates both fixed and stochastic effects and it allows the application of the Kalman filter and the Kalman smoother predictors. The parameters estimation is discussed and it is performed with both Gaussian maximum likelihood method and distribution-free estimators in a two-step procedure. The application of the Kalman smoother algorithm allows obtaining predictions of the structural components. The proposed approach is illustrated in the discrimination of the water monitoring sites using the monthly dissolved oxygen concentration dataset between January 2002 and May 2013, in the hydrological basin of the river Vouga, in Portugal. The water monitoring sites are discriminated through the structural components by a hierarchical agglomerative clustering procedure.

WAVELET-BASED METHODS

CG003 **Room Sala 7** **Chair: Ying Sun**

CC0242: **Using bandwidths as scales in multiscale localpolynomial decompositions**
Presenter: **Mohamed Amghar**, Universite Libre de Bruxelles, Belgium
Co-authors: Maarten Jansen

The choice of the bandwidths in a multiscale local polynomial data transform is discussed. The transform adopts the local polynomial smoothing paradigm for the construction of a multiresolution data decomposition, much like a wavelet transform or a Laplacian pyramid. The bandwidths depend on the resolution level, defining for each level the scale of the coefficients. As a result, the scale is not necessarily dyadic as in a discrete wavelet transform, nor is it grid dependent as in second generation wavelet transform. Unlike in a uniscale local polynomial smoothing scheme, the bandwidth in a multiscale data transform is not optimised for data processing, i.e. smoothing, but rather for data transformation. The bandwidth at each level should be chosen in a way that it makes the representation after transformation as suitable as possible for subsequent, non-linear processing. The objective is to find bandwidths that lead to optimal multiscale decomposition, in the sense that the resulting decomposition is as easy to work with. In particular we optimize the bandwidths with respect to sparsity of the decomposition on one hand, and the noise reduction of the decomposition through an orthogonal prefilters on the other hand.

CC0545: **Stochastic simulation models for multi-site nonstationary time series using wavelets**
Presenter: **Ying Sun**, KAUST, Saudi Arabia

Many meteorological variables exhibit nonstationarity in time. In particular, the time series appears to have modulated oscillations that may correspond to the recurrent but still changing set of climate conditions. A multi-scale stochastic simulation model is developed using wavelet decomposition. The multiresolution and localization properties of wavelets make them favorable for reproducing different time-varying local features of processes at different time scales. Specifically, for a given time scale, we propose to use a stochastic model based on evolving periodic functions to model the wavelet coefficients that vary in a periodic fashion, and both periodicity and amplitude are allowed to change over time. We apply this approach to analyze the monthly southern oscillation index from 1876 to 2015, and show that the proposed simulation model can successfully reproduce the interdecadal fluctuations, which have the effect of modulating the amplitude and frequency of occurrence of El Nino events. Multi-site simulation models are also developed with an application to a minute-by-minute meteorological dataset collected at multiple monitoring locations from the Atmospheric Radiation Measurement program.

CC0571: **Minimal support tight wavelet frames in a probabilistic MRI denoising method**
Presenter: **Sergio Villullas Merino**, University of Valladolid, Spain

Human body heat emission and others external causes can interfere in magnetic resonance image acquisition and produce noise. In this kind of images, the real Rician noise can be considered as Gaussian noise in high snr regions, and its wavelet frame coefficients can be approximately modeled by a Gaussian distribution. Noiseless magnetic resonance images can be modeled, in the wavelet frame domain, by a generalized Gaussian distribution with different fixed values of the parameter β depending of the scale of the wavelet frame descomposition (general or multirresolution). The image denoising method proposed performs a shrinkage of wavelet frame coefficients based on the conditioned probability of being noise or detail. The parameters involved in this filtering approach are calculated by means of the expectation maximization method, which avoids the need to use an estimator of noise variance. The efficiency of the proposed filter is studied and compared with other important filtering techniques, such as Nowak's, Donoho-Johnstone's, Awate-Whitaker's and non-local means filters, in different 2-dimensional images.

CC0260: **Disaggregated electricity forecasting using wavelet-based clustering of individual consumers**
Presenter: **Jairo Cugliari**, Univ Lyon - Ezus, France
Co-authors: Yannig Goude, Jean-Michel Poggi

Electricity load forecasting is crucial for utilities for production planning as well as marketing offers. Recently, the increasing deployment of smart grids infrastructure requires the development of more flexible data driven forecasting methods adapting quite automatically to new data sets. We propose to build clustering tools useful for forecasting the load consumption. The idea is to disaggregate the global signal in such a way that the sum of disaggregated forecasts significantly improves the prediction of the whole global signal. The strategy is in three steps: first we cluster curves defining super-consumers, then we build a hierarchy of partitions within which the best one is finally selected with respect to a disaggregated forecast criterion. The proposed strategy is applied to a dataset of individual consumers from the French electricity provider EDF. A substantial gain of 16% in forecast accuracy comparing to the 1 cluster approach is provided by disaggregation while preserving meaningful classes of consumers.

ALGORITHMS AND COMPUTATIONAL METHODS

CC063 **Room Sala 5** **Chair: Gerard Biau**

CC0394: **Expanded alternating optimization for matrix factorization and penalized regression**
Presenter: **William James Murdoch**, University of California at Berkeley, United States
Co-authors: Mu Zhu

We propose a general technique for improving alternating optimization (AO) of nonconvex functions. Starting from the solution given by AO, we conduct another sequence of searches over subspaces that are both meaningful to the optimization problem at hand and different from those

used by AO. To demonstrate the utility of our approach, we apply it to the matrix factorization (MF) algorithm for recommender systems and the coordinate descent algorithm for penalized regression (PR), and show meaningful improvements using both real-world (for MF) and simulated (for PR) data sets. Moreover, we demonstrate for MF that, by constructing search spaces customized to the given data set, we can significantly increase the convergence rate of our technique.

CC0154: **Semiparametric estimation of mixed analysis of covariance model**
Presenter: **Virgelio Alao**, Visayas State University, Philippines
Co-authors: Erniel Barrios, Joseph Ryan Lansangan

A semiparametric mixed analysis of covariance model is postulated and estimated using the two procedures: first, based on an embedded restricted maximum likelihood and nonparametric regression (smoothing splines) estimation into the backfitting framework; and second, infusing bootstrap into the first procedure. The heterogeneous effect of covariates across the groups is postulated to affect the response through a nonparametric function to mitigate overparameterization. Using simulation studies, we exhibited the capability of the postulated model (and estimation procedures) in increasing predictive ability and stabilizing variance components estimates even for small sample size and with minimal covariate effect, and regardless of whether the model is correctly specified or there is misspecification error.

CC0514: **Evaluation of similarity for contexts on association rule based extraction**
Presenter: **Ken Nittono**, Hosei University, Japan

The text mining approaches aiming to extract expressions which have particular features and gather systematically the resulting parts of documents have been increasing its importance in recent years. The contexts which are regarded as particular expressions are represented as combinations of terms and association rules in text mining methods are utilized for the extraction. The utilization of association rules enables to find essential combinations of terms valued throughout the large-sized target documents. The combinations of terms extracted by the association rules imply the pointers to the specific parts of original documents. Latent semantic analysis is applied in order to make this analyzing model have a relationship between combinations of terms and contests. Similarities between combinations of terms and the contexts are measured on a concept space generated by latent semantic analysis and it is decided which contexts throughout the whole documents have particular meanings. Herein, conditions such as influence of dimensionality of concept space on similarity and composition of cosine values as a similarity measure are evaluated. Collection of contexts which have significant similarity leads to generation of abstracted documents and, furthermore, accumulation of them enables applying to constructing a text database which is reusable as knowledge, for instance.

CC0586: **The coreset variational Bayes (CVB) algorithm for mixture analysis**
Presenter: **Clare McGrory**, University of Queensland, Australia

The pressing need for improved methods for analysing and coping with big data has opened up a new area of research for statisticians. Image analysis is an area where there is typically a very large number of data points to be processed per image, and often multiple images are captured over time. These issues make it challenging to design methodology that is reliable and yet still efficient enough to be of practical use. One promising emerging approach for this problem is to reduce the amount of data that actually has to be processed by extracting what we call coresets from the full dataset; analysis is then based on the coreset rather than the whole dataset. Coresets are representative subsamples of data that are carefully selected via an adaptive sampling approach. We propose a new approach called coreset variational Bayes (CVB) for mixture modelling; this is an algorithm which can perform a variational Bayes analysis of a dataset based on just an extracted coreset of the data. We apply our algorithm to weed image analysis.

STATISTICAL COMPUTING

CG105 **Room Sala 4** **Chair: Sugnet Lubbe**

CC0445: **Consequences of combining Fisher's optimal scores with biadditive models**
Presenter: **Niel Le Roux**, Stellenbosch University, South Africa
Co-authors: John Gower, Sugnet Lubbe

Fisher's optimal scores quantify a dependent categorical variable of a two-way table. The quantifications maximising the additive part of a biadditive model are found as the eigenvector associated with the largest eigenvalue satisfying an eigenequation. The non-additive part of the model can then be inspected by constructing a biadditive biplot based on a singular value decomposition (svd) of the residuals. However, this biplot is not optimal. It can be shown that the non-additive part of the model can be optimally represented using the eigenvector that is associated with the smallest eigenvalue in the above eigenequation suggesting the construction of an improved biplot for representing the non-additive part of the biadditive model. An important question is what about the intermediate solutions to the eigenvalue problem. The quantifications found from all possible solutions can be arranged as the columns of a matrix leading to a Long form of the problem parallel to the Short form version in terms of the original two-way table. Thus two routes for analysis become available: finding all solutions of the eigenequation arising from the Short form or an svd of the Long form. Links between these two routes are discussed and some extensions to existing biadditive biplots are proposed.

CC0216: **Estimation of reliability with semi-parametric frailty modeling of degradation**
Presenter: **Prajamitra Bhuyan**, Indian Statistical Institute, India
Co-authors: Debasis Sengupta

In many real life scenarios, stress accumulates over time and the system fails as soon as the accumulated stress or degradation equals or exceeds a critical threshold. For some devices, it is possible to obtain measurements of degradation over time, and these measurements may contain useful information about product reliability. We propose a semi-parametric random effect (frailty) model for degradation path, and a method of estimating this path as well as the reliability. Consistency of the estimator under general conditions is established. Simulation results show superiority of the performance of the proposed method over a parametric competitor. The method is illustrated through the analysis of a real data set.

CC0362: **Overflow analysis of multiple stacks running on the same memory**
Presenter: **Kamil Demirberk Unlu**, Ankara University and Middle East Technical University, Turkey
Co-authors: Ali Devin Sezer

A basic mathematical model used for the analysis of dynamic storage allocation is that of a constrained random walk X on the positive orthant $\mathbb{Z}_+^d$ with increments $\mathcal{V} = \{-e_i, +e_i, i = 1,2,3,...,d\}$, where $\{e_i, i = 1,2,3,...,d\}$ is the standard basis for $\mathbb{R}^d$, i.e.,$X_{k+1} = X_k + \pi(X_k, Y_k)$ where $\{Y_k\}$ is an i.i.d sequence taking values in $\mathcal{V}$ and $\pi(x,y) = y$ if $x+y \in \mathbb{Z}_+^d$ and 0 otherwise. This walk models d independent stacks randomly inserting and deleting elements in a jointly used memory of size n. Let $\tau_n \doteq \inf\left\{k : \sum_{i=1}^d X_i(k) = n\right\}$, the time when the joint buffer holding these stacks overflows. A much studied quantity in the analysis of this system is $\mathbb{E}\left[\max_{i=1}^d X_i(\tau_n)\right]$, i.e., the expected size of the longest stack at the time of buffer overflow. The goal is to extend the approximation of the expected size of the longest stack at the time of buffer overflow and the probability that a buffer overflow occurs before the system empties for the constrained random walk representing the stack model. To the best of our knowledge, the current literature on the subject mostly focuses to the case of $d = 2$, the goal is to compute approximations also for $d = 3$ and, if possible, for $d > 3$.

CC0485: **Monte Carlo studies when all models are wrong**
Presenter: **Mark De Rooij**, Leiden University, Netherlands

In statistics we often use Monte Carlo simulation studies to investigate properties of statistical models in different circumstances. Most of the time we adopt a true data generating model and study the effect of misspecification of errors, structural form, or other disturbances on the bias and variance of parameter estimates and their standard errors. However, in data analytic settings most statisticians agree that all models are wrong. Therefore, the above mentioned Monte Carlo studies have little utility for data analysis. We will discuss this problem in detail and show alternative ways of performing simulation studies with targeted models. These models do not assume that the true model can be attained but instead define a target of analysis. We will illustrate the methodology using linear regression and latent variable models.

Tuesday 23.08.2016 16:30 - 18:00 Parallel Session D – COMPSTAT

RECENT DEVELOPMENTS IN MIXTURE MODELS

CI083 **Room Sala Camara** **Chair: Francesca Greselin**

CO0263: **Linear regression models with finite mixtures of skew heavy-tailed errors**
Presenter: **Victor Hugo Lachos Davila**, UNICAMP, Brazil
Co-authors: Luis Enrique Benites Sanchez, Rocio Maehara Aliaga

The aim is to estimate regression models whose error terms follow a finite mixture of scale mixtures of skew-normal (SMSN) distributions, a rich class of distributions that contains the skew-normal, skew-t, skew-slash and skew-contaminated normal distributions as proper elements. This approach allows us to model data with great flexibility, accommodating simultaneously multimodality, skewness and heavy tails. We developed a simple EM-type algorithm to perform maximum likelihood (ML) inference of the parameters of the proposed model with closed form expressions for both E- and M-steps. Furthermore, the empirical information matrix is derived analytically to account for standard errors. The practical utility of the new method is illustrated with the analysis of a real dataset and several simulation studies. The proposed algorithm and methods are implemented in the R package FMsmsnReg.

CI0376: **Recent developments in model-based clustering of functional data**
Presenter: **Charles Bouveyron**, University Paris Descartes, France

Some recent advances in model-based clustering of functional data will be reviewed. Then, the focus will be on a recently proposed methodology. The motivation will be based on the analysis and comparison of several European BSSs to identify common operating patterns in BSSs and to propose practical solutions to avoid potential issues. Our approach relies on the identification of common patterns between and within systems. To this end, a model-based clustering method, called FunFEM, for time series (or more generally functional data) is developed. It is based on a functional mixture model that allows the clustering of the data in a discriminative functional subspace. This model presents the advantage in this context to be parsimonious and to allow the visualization of the clustered systems. Numerical experiments confirm the good behavior of FunFEM, particularly compared to state-of-the-art methods. The application of FunFEM to BSS data from JCDecaux and the Transport for London Initiative allows us to identify 10 general patterns, including pathological ones, and to propose practical improvement strategies based on the system comparison. The visualization of the clustered data within the discriminative subspace turns out to be particularly informative regarding the system efficiency. The proposed methodology is implemented in a package for the R software, named funFEM, which is available on the CRAN. The package also provides a subset of the data analyzed.

CI0385: **On mixture modelling with multivariate skew distributions**
Presenter: **Geoffrey McLachlan**, University of Queensland, Australia
Co-authors: Sharon Lee

In recent years, there has been increasing use of non-normal distributions in the modelling and analysis of heterogeneous data. Attention is focussed on the use of skew symmetric distributions with multivariate skewing functions that allow for the modelling of skewness in p arbitrary directions in the feature space, where p is the number of variables. In particular, various multivariate skew normal and skew t- distributions are considered corresponding to some commonly used characterizations in the literature. Parameter estimation for these distributions and mixtures of them can be obtained via the Expectation-Maximization (EM) algorithm. However, the E-step for such models typically involves the calculation of multidimensional integrals that are computationally expensive to evaluate. Some approaches are therefore considered to reduce the computation time required for the fitting of these models. In addition to methods that are directly applicable to single-threaded implementation, an approach is developed that utilizes the processing resources available from machines with multiple cores. An example on a real dataset will be given to illustrate the approach.

NONPARAMETRIC METHODS FOR ROC CURVES

CO008 **Room Sala 1** **Chair: Juan-Carlos Pardo-Fernandez**

CO0257: **Nonparametric and parametric confidence intervals for the Youden index and its associated cutoff point**
Presenter: **Benjamin Reiser**, University of Haifa, Israel
Co-authors: Leonidas Bantis, Christos T Nakas

The receiver operating characteristic (ROC) curve is commonly used to evaluate a continuous biomarker that attempts to discriminate between a healthy and a diseased population. The ROC curve is a plot of the sensitivity $\{sens(c)\}$ versus 1-specificity $\{1-spec(c)\}$ over all possible threshold values c of the marker. To evaluate the discriminatory ability of a marker it is common to summarize the information of the ROC curve into a single global value or index. Although the area under the ROC curve is the most frequently used global index of diagnostic accuracy the maximum of the Youden Index, being defined by maximizing over all possible threshold values as $J = \max\{sens(c) + spec(c) - 1\}$, is also often used. J is equivalent to the Kolmogorov-Smirnov distance between the two populations. In practice, clinicians are often interested in determining a cutoff for classification purposes. Frequently the optimal cutoff value c^* is chosen as the value of c for which J is maximized. In the applied literature confidence intervals for J and c^* are typically ignored. We provide new nonparametric kernel density-based and parametric delta method-based methods for constructing confidence intervals for both J and c^*. We compare our methods to currently available techniques through simulations and discuss some real examples.

CO0319: **Efficient confidence bands for ROC curves**
Presenter: **Sonia Perez-Fernandez**, University of Oviedo, Spain
Co-authors: Pablo Martinez-Camblor, Norberto Corral

The receiver operating characteristic (ROC) curve is a popular graphical method frequently used in order to study the diagnostic capacity of continuous (bio)markers. In spite of the existence of a huge number of papers devoted to both the theoretical and practical aspects of this topic, the construction of confidence bands has had little impact in the specialized literature. As far as the authors know, there are no R packages providing ROC curve confidence bands. We try to fill this gap studying and proposing a method to build confidence bands for both the usual and the general ROC curves. The behaviour of the proposed methodology is studied via Monte Carlo simulations and applied on real-world biomedical problems. In addition, an R function to compute the proposed and some previously existing methodologies is provided.

CO0297: **A comparative study of methods for testing the equality of two or more ROC curves**
Presenter: **Aris Fanjul Hevia**, Universidad de Santiago de Compostela, Spain
Co-authors: Wenceslao Gonzalez-Manteiga

The problem of comparing the accuracy of diagnostic tests is usually carried out through the comparison of the corresponding ROC curves. This matter has been approached from different perspectives. Usually, ROC curves are compared from their respective AUCs, but in the case where there is no uniform dominance between the involved curves other procedures are preferred, such as the ones based on the distance of two empirical quantil processes, or the ones that take advantage of the fact that the ROC curves can be viewed as cumulative distribution functions. Although the asymptotic distributions of the statistics behind these methods are, in general, known, resampling plans are considered. With the purpose of

comparing the performance of these different approaches, a simulation study is carried out. Additionally, some of these methods are extended to the case where covariates are available.

CO0318: **Smooth time-dependent ROC curve estimators**
Presenter: **Juan-Carlos Pardo-Fernandez**, Universidade de Vigo, Spain
Co-authors: Pablo Martinez-Camblor

The receiver operating characteristic (ROC) curve is a popular graphical method often used to study the diagnostic capacity of continuous (bio)markers. When the considered outcome is a time-dependent variable, two main extensions have been proposed: the cumulative/dynamic ROC curve and the incident/dynamic ROC curve. In both cases, the main problem for developing appropriate estimators is the estimation of the joint distribution of the time-to-event variable and the marker. As usual, different approximations lead to different estimators. We explore the use of a bivariate kernel density estimator which accounts for censored observations in the sample and produces smooth estimators of the time-dependent ROC curves. The performance of the resulting cumulative/dynamic and incident/dynamic ROC curves is studied by means of Monte Carlo simulations. Additionally, the influence of the choice of the required bandwidth is explored. The obtained results suggest that the smooth estimators provide good approximations, specially when the area under the ROC curve is not too large.

COPULA MODELLING IN NONLINEAR TIME SERIES

CO089 **Room Sala 5** **Chair: Richard Gerlach**

CO0280: **Inversion copulas from nonlinear state space models**
Presenter: **Michael Smith**, University of Melbourne, Australia
Co-authors: Worapree Ole Maneesoonthorn

While copulas constructed from inverting latent elliptical, or skew-elliptical, distributions are popular, they can be inadequate models of serial dependence in time series. As an alternative, we propose an approach to construct copulas from the inversion of latent nonlinear state space models. This allows for new time series copula models that have the same serial dependence structure as a state space model, yet have an arbitrary marginal distribution something that is difficult to achieve using other time series models. We examine the time series properties of the copula models, outline measures of serial dependence, and show how to use Bayesian methods to estimate the models. To illustrate the breadth of new copulas that can be constructed using our approach, we consider example latent state space models. We use the resulting three inversion copulas to model and forecast quarterly U.S. inflation data. We show how combining the serial dependence structure of the state space models, with flexible asymmetric and heavy-tailed margins, improves the accuracy of the fit substantially.

CC0259: **Dynamic quantile function modelling**
Presenter: **Richard Gerlach**, University of Sydney, Australia

Modelling the time-varying distributions of financial returns has been an interest to many authors in recent decades. The increasing availability of high-frequency data has presented new challenges, namely, effectively making use of the noisy information contained in the intra-daily observations at a reasonable computational cost. Borrowing ideas from symbolic data analysis (SDA), we consider data representations beyond scalars and vectors. Specifically, we consider a quantile function as an observation, and propose a class of dynamic models for quantile-function-valued time series. Direct modelling of quantile functions can be more convenient in applications where the quantity of interest is a quantile. We present a method whereby a likelihood function can be defined for quantile function-valued data, and develop an MCMC algorithm for simulating from the posterior distribution. In the empirical study, we model the time series of quantile functions of high frequency financial returns, and demostrate the usefulness of our method by forecasting one-step-ahead the extreme quantiles of intra-daily returns. Furthermore, through a simple empirical scaling rule, we are able to forecast one-step-ahead the Value-at-Risk of daily returns.

CC0470: **Vine based modeling of multivariate volatility time-series**
Presenter: **Nicole Barthel**, Technische Universitaet Muenchen, Germany
Co-authors: Claudia Czado, Yarema Okhrin

Studying realized volatility based on high-frequency data is of particular interest in asset pricing, portfolio management and evaluation of risk. We propose an approach for dynamic modeling and forecasting of realized correlation matrices that allows for model parsimony and automatically guarantees positive definiteness of the forecast. We use the one-to-one relationship between a correlation matrix and its associated set of partial correlations corresponding to any regular vine specification. Being algebraically independent the partial correlations of the vine do not need to satisfy any algebraic constraint such as positive definiteness. We present several selection methods to choose, among the large number of vines, the vine structure best capturing the characteristics of the multivariate time-series of the correlation parameters. The individual partial correlation time-series are studied using ARFIMA and HAR models to account for long-memory behavior. The dependence between assets is flexibly modeled using vine copulas that allow for nonlinearity and asymmetric tail behavior. Correlation point forecasts are obtained as nonlinear transformation of the forecasts for the partial correlation vine. The usefulness of the methodology is investigated in a one-day ahead forecasting framework comparing existing methods based on a model confidence set approach.

CC0568: **Copula-MGARCH with dynamic conditional covariance decompositions**
Presenter: **Fabian Raters**, University of Goettingen, Germany
Co-authors: Helmut Herwartz

The Copula-MGARCH (C-MGARCH) model incorporates standardized copula distributed innovations in MGARCH models. Recent literature suggests that a static continuous decomposition of the standardized innovations' covariance matrix enhances the model's flexibility. A further generalization of the C-MGARCH approach is identified by means of dynamic decompositions of the conditional covariances. These decompositions are assumed to depend on a latent univariate autoregressive process and, considering bivariate BEKK, CCC, and DCC models, dynamically rotate the standardized innovations. Contributing to the recent literature, we suggest a stepwise estimation procedure of the models' parameters by means of Maximum Likelihood and particle filtering. The performance of our extension applied to the aforementioned models is evaluated in a comprehensive study. Empirically, we conduct an application to the log-differences of daily exchange rates by means of in-sample information criteria and ex-ante portfolio Value-at-Risk coverage tests. Our approach contributes to the modeling of nonlinear dependencies between multivariate time series of log-differences in volatility models.

NEW ADVANCES IN MULTISET AND MULTIWAY DATA ANALYSIS II

CO110 **Room Sala 3** **Chair: Eva Ceulemans**

CO0283: **The finite iterative method for the computation of the correlation matrix implied by a structural recursive model**
Presenter: **Mohamed Hanafi**, ONIRIS, France
Co-authors: Zouhair El hadri, El Kettani Yousfi

Structural Equation Modeling is a multivariate powerful technique used for analyzing causal relationship between hypothetical constructs. The computation of the covariance (or correlation) matrix implied by the model is a crucial step in this approach. Usually, the method used for this task is called Jorskog method. An alternative and a new method called the Finite Iterative Method (FIM) is introduced. Recently, this method has been published for models without latent variables (Path Analysis). Its extension to the structural equation models with latent variables is considered.

The efficiency of the Finite Iterative Method compared with the known one will be discussed. How the new method allows simple procedures for the estimation of the parameters in Structural Equation Modeling is suggested.

CC0267: **Searching components with simple structure in simultaneous component analysis: Blockwise simplimax rotation**
Presenter: **Henk Kiers**, University of Groningen, Netherlands
Co-authors: Marieke Timmerman, Eva Ceulemans
Simultaneous component analysis (SCA) is a fruitful approach to disclose the structure underlying data stemming from multiple sources on the same objects. This kind of data can be organized in blocks. To identify which component relates to all, and which to some sources, the block structure in the data should be taken into account. We propose a new rotation criterion, Blockwise Simplimax, that aims at block simplicity of the loadings, implying that for some components all variables in a block have a zero loading. We also present an associated model selection criterion, to aid in selecting the required degree of simplicity for the data at hand. An extensive simulation study is conducted to evaluate the performance of Blockwise Simplimax and the associated model selection criterion, and to compare it with a sparse competitor, namely Sparse group SCA. In the conditions considered Blockwise Simplimax performed reasonably well, and either performed equally well as, or clearly outperformed Sparse group SCA. The model selection criterion performed well in simple conditions. The usefulness of Blockwise Simplimax and Sparse group SCA is illustrated using sensory profiling data regarding different cheeses.

CO0344: **A sparse common component approach with selection of relevant clusters of variables**
Presenter: **Katrijn Van Deun**, Tilburg University, Netherlands
Social science research has entered the era of big data: Many detailed measurements are taken and multiple sources of information are used to unravel complex multivariate relations. For example, in studying obesity as the outcome of environmental and genetic influences, researchers increasingly collect survey, dietary, biomarker and genetic data from the same individuals. Although linked more-variables-than-samples (called high-dimensional) multi-source data form an extremely rich resource for research, extracting meaningful and integrated information is challenging and not appropriately addressed by current statistical methods. A first problem is that relevant information is hidden in a bulk of irrelevant variables with a high risk of finding incidental associations. Second, the sources are often very heterogeneous, which may obscure apparent links between the shared mechanisms. Hence, a statistical framework is needed to (i) select the relevant groups of variables within each source and (ii) link them throughout data sources. To address these issues, we present an extension of sparse simultaneous component analysis to a component method that 1) finds the common components and 2) that selects relevant clusters of variables.

CO0330: **Simultaneous principal and network components models for integration of multiset data**
Presenter: **Pia Tio**, University of Amsterdam and Tilburg University, Netherlands
With increasingly more sophisticated instruments and growing interdisciplinary cooperation, more and more huge datasets are being gathered that contain, for the same individuals, detailed information coming from multiple data sources. Given their heterogeneity and high-dimensionality, relevant information is hidden in a bulk of irrelevant variables and there is a high risk of finding (spurious) associations by coincidence. To take full advantage of the availability of multiset data, mature data integration techniques that yield interpretable results derived from a massive amount of data are needed. While data reduction techniques can deal with some of these obstacles, the interpretation of their resulting components/factors is difficult. We propose that the combined efforts of component and network analysis can deal with these obstacles simultaneously. Component analysis techniques such as sparse simultaneous component analysis are especially useful to extract subtle common processes amongst dominant source-specific variation while at the same time reducing the dimensionality of the data. Network analysis provides means to investigate unique processes amongst mutually interacting components/variables within a source without having to resort to latent variables. We present results from simulations showing the performance of the combination of the component and network analyses.

REGULARIZATION

CO019 **Room Sala 4** **Chair: Andreas Alfons**

CC0449: **SparseStep: Approximating the counting norm for sparse regularization**
Presenter: **Gertjan van den Burg**, Erasmus University Rotterdam, Netherlands
Co-authors: Andreas Alfons, Patrick Groenen
The SparseStep algorithm is presented for the estimation of a sparse parameter vector in the linear regression problem. The algorithm works by adding an approximation of the exact counting norm as a constraint on the model parameters, and iteratively strengthening this approximation to arrive at a sparse solution. Theoretical analysis of the penalty function shows that the estimator yields unbiased estimates of the parameter vector. An iterative majorization algorithm is derived which has a straightforward implementation reminiscent of ridge regression. In addition, the SparseStep algorithm is compared with similar methods through a rigorous simulation study which shows it often outperforms existing methods in both model fit and prediction accuracy.

CC0208: **Regularized clusterwise multiblock regression**
Presenter: **Stephanie Bougeard**, ANSES, France
Co-authors: Ndeye Niang, Gilbert Saporta
Regression coefficients are usually estimated under the assumption that observations come from a single and homogeneous population. However in many applications, this assumption is not true and the overall model is not efficient to recover the specificities of potential cluster models. When variables are in addition structured into a dependent block of variables and several blocks of explanatory ones, we propose a new method called regularized clusterwise multiblock regression. The aims of this method are to find out simultaneously thanks to a single criterion: a data reduction of explanatory variables through components that can be intermediate between the ones from multiblock PLS and multiblock Redundancy Analysis, a partition of the observations into several clusters and the corresponding cluster multiblock regression coefficients. The three unknown parameters of this criterion, namely the regularization parameter which aims at stabilize the inversion of the block variance-covariance matrices, the number of components and the number of clusters, are all defined such as to minimize the prediction error on the basis of a ten-fold cross-validation. A simulation study is carried out to assess the performance of the method and an empirical application in the field of consumer satisfaction is provided which illustrates the usefulness of the method.

CC0390: **Extension to mixed models of the supervised component-based generalised linear regression**
Presenter: **Jocelyn Chauvet**, University of Montpellier, France
Co-authors: Catherine Trottier, Xavier Bry, Frederic Mortier
The component-based regularisation of a multivariate Generalized Linear Mixed Model (GLMM) is adressed. A set of random responses Y is modelled by a GLMM, using a set X of explanatory variables, a set T of additional covariates, and random effects used to introduce the dependence between statistical units. Variables in X are assumed many and redundant, so that regression demands regularisation. By contrast, variables in T are assumed few and selected so as to require no regularisation. Regularisation is performed building an appropriate number of orthogonal components that both contribute to model Y and capture relevant structural information in X. To estimate the model, we propose to maximise a criterion specific to the Supervised Component-based Generalised Linear Regression (SCGLR) within an adaptation of Schall's algorithm. This extension of SCGLR is tested on both simulated and real data, and compared to Ridge- and Lasso-based regularisations.

CC0392: **Supervised-component based Cox regression**
Presenter: **Xavier Bry**, Universite Montpellier, France
Co-authors: Theo Simac, Thomas Verron
In survival analysis with high-dimensional regressors, the Cox Proportional Hazard Model (CPHM) encounters instability problems owing to regressor-collinearity. Its regularisation is therefore needed. Two family of methods currently perform regularisation of regression models: penalty-based methods such as Ridge and Lasso, and component-based models such as PLS-type regression models. Only the latter enable exploratory analysis of predictive structures by making components lean on the regressors' strong correlation structures. We propose a new way to take into account the structural relevance of components within the estimation procedure of the Cox Model. Our algorithm, called SCCoxR (for Supervised-Component based Cox Regression), is tested on simulated data, and then, applied to life-history data of HIV-positive Thai subjects, in order to model the age at disclosure of their serologic status.

GRAPHICAL MODELS AND NETWORKS

CG044 **Room Sala 7** **Chair: Lourens Waldorp**

CC0566: **State space models for the NGS pipeline**
Presenter: **Karin Dorman**, Iowa State University, United States
The age-old wisdom "garbage in, garbage out" underscores any analysis using next-generation sequencing (NGS) data. Pipeline components are concatenated serially with only minimal transmission of uncertainties and information. For example, base callers rarely utilize information about the underlying genome sequence, whereas error correction methods seldom utilize the error properties of sequencing. We demonstrate that integrated, probabilistic approaches that combine steps in the pipeline perform better than sequential analysis. Others have improved pipeline operations by borrowing information from alignment to known reference genome(s). Our combined approach specifically capitalizes on genome information, but without use of a known reference genome to avoid biasing against the unknown. We use a Hidden Markov Model on a sparse de Bruijn graph, where the transitions model genetic content and the emissions model observable data. The combined probabilistic approach removes more errors and more accurately transmits information through the pipeline.

CC0184: **Estimating joint sparse graphical models using fMRI brain imaging data with different coarseness levels**
Presenter: **Eugen Pircalabelu**, KU Leuven, Belgium
Brain networks are estimated from fMRI datasets that do not all contain measurements on the same set of regions. For certain datasets, some of the regions have been split in smaller subregions, while others have not been split. This gives rise to mixed scale measurements and the purpose is to estimate sparse undirected graphical models. To overcome the problem of mixed coarseness levels, the expand and the reduce algebraic operators are used. To estimate sparse undirected graphs, the ADMM algorithm is used and to ensure similarity of graphs across coarseness levels, the procedure uses the fused and group lasso penalties for certain block submatrices and a lasso penalty for the remaining submatrices. The method estimates edges for each subject and coarseness level, referred to as within level edges, and identifies possible connections between a large region and its subregions, referred to as between level edges which offer insight into whether a certain large region is constructed by aggregating homogeneous or heterogeneous parts of the brain. It also avoids the tedious task of selecting one coarseness level for carrying out the analysis and produces interpretable results at all available levels. Empirical and theoretical evaluations illustrate the usefulness of the method.

CC0509: **Using chain graph models for structural inference for multi-level data with an application to linguistic data**
Presenter: **Craig Alexander**, University of Glasgow, United Kingdom
Co-authors: Ludger Evers, Tereza Neocleous, Jane Stuart-Smith
Graphical models provide a visualisation of the conditional dependence structure between variables, making them an attractive tool for inference. We discuss the use of a chain graph model as an alternative to a standard multi-level regression analysis with a multivariate response. The improved readability of the model output makes this an appealing alternative for those not with a strong statistical background. The chain graph can be inferred from three parts. The dependency structure of the covariates can be modelled independently using standard methods for structural inference in graphical models such as log-linear models in the case of categorical covariates. The directed edges between the explanatory and response variables and the undirected edges between the response variables are jointly inferred using a multivariate Bayesian multi-level model in which the precision matrix of residuals and random effects is assumed to conform to a undirected graphical model, which is to be inferred. We present an application of this model using linguistic data obtained from the Sounds of the City corpus which consists of a real time corpus of Glaswegian speech. From the data, we look to recover the underlying chain graph model detailing which factors affect vowel quality.

ROBUST METHODS IN REGRESSION PROBLEMS

CG020 **Room Sala 2** **Chair: Luis Angel Garcia-Escudero**

CC0212: **Robust inference in a heteroskedastic multilevel linear model with structural change**
Presenter: **Ronrick Da-ano**, University of the Philippines-Diliman, Philippines
Co-authors: Erniel Barrios, Joseph Ryan Lansangan
A heteroskedastic multilevel model with cross-sectional interactions in higher and lower levels with structural change is estimated by a hybrid procedure of the forward search algorithm preceding bootstrap method. The simulation study exhibits the ability of the hybrid procedure to produce narrower confidence interval even when there is model misspecification error and structural change. Moreover, it has a comparable predictive ability with the classical restricted maximum likelihood (REML) estimation. However, the hybrid method yield estimates of the parameters with lower bias relative to REML. The hybrid of forward search and bootstrap method can further robustify estimates of fixed and random coefficients under various levels of interclass correlation or in the presence of structural change in a multilevel model.

CC0352: **Coping with level and character of contamination by SW-estimator**
Presenter: **Jan Amos Visek**, Charles University in Prague, Czech Republic
A new estimator inheriting (hopefully) pros and removing cons of the *S-estimator* and the *least weighted squares* is proposed. It allows for both, a weight function w as well as for an objective function ρ which need not be bounded. The conditions for consistency and asymptotic representation are recalled. The combination of the weight function with the objective function allows to accommodate the estimator to the level and character of contamination. The results of simulation studies confirm it.

CC0500: **L-moments in linear regression models**
Presenter: **Martin Schindler**, Technical University of Liberec, Czech Republic
Co-authors: Jan Picek
The L-moments are analogues of the conventional moments and have similar interpretations. They are calculated using linear combinations of the expectation of ordered data. L-moments have been shown to be special case of L-estimators. We propose a generalization of L-moments in the linear regression model based on regression quantiles as special L-estimator. The properties of extended L-moments are illustrated on simulated data.

CC0368: **Robust quantile regression in R**

Presenter: **Christian Galarza**, State University of Campinas, Brazil
Co-authors: Luis Enrique Benites Sanchez, Victor Hugo Lachos Davila

It is well known that the widely popular mean regression model could be inadequate if the probability distribution of the observed responses do not follow a symmetric distribution. To deal with this situation, the quantile regression turns to be a more robust alternative for accommodating outliers and the misspecification of the error distribution since it characterizes the entire conditional distribution of the outcome variable. A likelihood-based approach is presented for the estimation of the regression quantiles based on a new family of skewed distributions. This family includes the skewed version of Normal, Student-t, Laplace, contaminated Normal and slash distribution, all with the zero quantile property for the error term, and with a convenient and novel stochastic representation which facilitates the implementation of the EM algorithm for maximum-likelihood estimation of the *p*th quantile regression parameters. We evaluate the performance of the proposed EM algorithm and the asymptotic properties of the maximum-likelihood estimates through empirical experiments and application to a real life dataset. The algorithm is implemented in the R package `lqr()`, providing full estimation and inference for the parameters as well as simulation envelopes plots useful for assessing the goodness-of-fit.

Wednesday 24.08.2016 09:00 - 10:30 Parallel Session F – COMPSTAT

ADVANCES IN RANDOM FORESTS

CI075 **Room Sala Camara** **Chair: Jean-Michel Poggi**

CI0155: **Causal inference with random forests**
Presenter: **Stefan Wager**, Stanford University, United States

Many scientific and engineering challenges -ranging from personalized medicine to customized marketing recommendations- require an understanding of treatment heterogeneity. We develop a non-parametric causal forest for estimating heterogeneous treatment effects that extends Breiman's widely used random forest algorithm. Given a potential outcomes framework with unconfoundedness, we show that causal forests are pointwise consistent for the true treatment effect, and have an asymptotically Gaussian and centered sampling distribution. We also propose a practical estimator for the asymptotic variance of causal forests. In both simulations and an empirical application, we find causal forests to be substantially more powerful than classical methods based on nearest-neighbor matching, especially as the number of covariates increases. Our theoretical results rely on a generic asymptotic normality theory for a large family of random forest algorithms. To our knowledge, this is the first set of results that allows any type of random forest, including classification and regression forests, to be used for formally valid statistical inference.

CI0173: **Tests for nonparametric interactions using random forests**
Presenter: **Giles Hooker**, Cornell University, United States
Co-authors: Lucas Mentch

While statistical learning methods have proved powerful tools for predictive modeling, the black-box nature of the models they produce can severely limit their interpretability and the ability to conduct formal inference. However, the natural structure of ensemble learners like bagged trees and random forests has been shown to admit desirable asymptotic properties when base learners are built with proper subsamples. We demonstrate that by defining an appropriate grid structure on the covariate space, we may carry out formal hypothesis tests for both variable importance and underlying additive model structure. To our knowledge, these tests represent the first statistical tools for investigating the underlying regression structure in a context such as random forests. We develop notions of total and partial additivity and further demonstrate that testing can be carried out at no additional computational cost by estimating the variance within the process of constructing the ensemble. Furthermore, we propose a novel extension of these testing procedures utilizing random projections in order to allow for computationally efficient testing procedures that retain high power even when the grid size is much larger than that of the training set.

CI0274: **Random forests variable importances: Towards a better understanding and large-scale feature selection**
Presenter: **Pierre Geurts**, University of Liege, Belgium
Co-authors: Antonio Sutera, Gilles Louppe, Celia Chatel, Louis Wehenkel

One of the most practically useful features of random forests is the possibility to derive from the ensemble of trees an importance score for each input variable that assesses its relevance for predicting the output. These importance scores have been successfully applied on many problems but they are still not well understood theoretically. Recent works towards a better understanding, and a better exploitation, of the mean decrease impurity (MDI) measure will be discussed. First, a theoretical analysis of this measure in asymptotic sample and ensemble size conditions will be presented. Main results include a characterization of the conditions under which this measure is consistent with respect to a common definition of variable relevance. Then, motivated by very high dimensional problems, MDI importances derived from finite tree ensembles will be analysed under the constraint that each tree can be built only from a subset of variables of fixed size. In this setting, a sequential variable sampling mechanism is proposed and compared with uniform sampling. When used for the identification of all relevant variables, importance scores obtained using this sampling mechanism are shown, theoretically and empirically, to significantly improve convergence speed in several conditions with respect to uniform sampling.

ARS-IASC SESSION III: NATURE-INSPIRED ALGORITHMS AND MULTIPLE RESPONSE OPTIMIZATION

CO045 **Room Sala 5** **Chair: Frederick Kin Hing Phoa**

CO0197: **On algorithms for generating multiple optimal experimental designs**
Presenter: **Yongtao Cao**, Indiana University of Pennsylvania, United States

The development of high performance algorithms for generating multiple optimal experimental designs is discussed. A new Pareto-based coordinate exchange algorithm for populating or approximating the true Pareto front for multi-criteria optimal experimental design problems will be emphasized. This heuristic combines an elitist-like operator inspired by evolutionary multi-objective optimization algorithms with a coordinate exchange operator that is commonly used to construct optimal designs. Benchmarking results from two to four dimensional and from screening design to split-plot design examples demonstrate that the proposed hybrid algorithm can generate highly reliable Pareto fronts with less computational effort than existing procedures in the statistics literature. The proposed algorithm also utilizes a multi-start operator, which makes it readily parallelizable for high performance computing infrastructures.

CO0239: **A multi-objective implementation in swarm intelligence with applications in designs of computer experiments**
Presenter: **Frederick Kin Hing Phoa**, Academia Sinica, Taiwan
Co-authors: Livia Lin-Hsuan Chang

As technology has advanced nowadays, it is common for a system to be optimized under more than one objective function, which leads to inconclusive results if two contradictory objectives exist. Traditional approaches suggest a simple aggregation of multiple objectives into one via a linear combination, but it is hard to justify the weights quantitatively. A systematic therapy to multiple objective optimization problem is proposed: (1) When the importance of criteria are known in prior, a sequential optimization is conducted; and (2) When the criteria are known to be equally important, a simple aggregated objective function with equal weights is suggested. The swarm intelligence based (SIB) method is extended for multiple objectives, namely Multiple Objective Swarm Intelligence Based (MOSIB) method. This method is then applied to the search of optimal designs of computer experiments, Latin hypercibe designs (LHDs), under several common criteria. Numerical studies show that the MOSIB method successfully generates a new series of optimal LHDs that possess better design properties than those suggested in the literature.

CO0241: **Nature-inspired meta-heuristic algorithms for generating optimal experimental designs**
Presenter: **Weng Kee Wong**, UCLA, United States

Nature-inspired meta-heuristic algorithms are increasingly studied and used in computer science and engineering disciplines to solve high-dimensional complex optimization problems in the real world. It appears relatively few of these algorithms are used in mainstream statistics even though they are simple to implement, very flexible and frequently able to find an optimal or a nearly optimal solution quickly. These general optimization methods usually do not require any assumption on the function to be optimized and the user only needs to input a few tuning parameters. It is given an overview of such algorithms, demonstrate the usefulness of some of these algorithms for finding different types of optimal designs for nonlinear models, suggest what to do if they don't seem to work and ascertain their overall potential.

CO0312: **Solving large scale penalized l_0-norm regression problems via parallel proximal algorithms**
Presenter: **Tso-Jung Yen**, Academia Sinica, Taiwan

An algorithm is developed for solving a regression estimation problem involving a structured l_0-norm penalty function. This algorithm incorporates the ideas of the proximal gradient method and iterative hard-thresholding. It decomposes the computational task into several sub-tasks that can be carried out separately in parallel. It obtains updates for parameters by using a closed form representation for the proximal operator of the structured l_0-norm penalty function. It is scalable in terms of sample size or the number of parameters, and is able to be implemented under a data parallelism framework. We demonstrate performance of the algorithm by conducting several simulation studies.

ADVANCES IN COMPUTATIONAL STATISTICS AND STATISTICAL MODELLING I
CO006 **Room Sala 1** **Chair: Pedro Jodra-Esteban**

CO0178: **The power Muth distribution**
Presenter: **Pedro Jodra-Esteban**, Universidad de Zaragoza, Spain
Co-authors: Hector W Gomez, Maria Dolores Jimenez-Gamero

Muth defined a continuous probability distribution with application in reliability theory. A new distribution is derived from the Muth distribution. Some statistical properties of the model are studied, such as the computation of the moments, generation by computer of pseudo-random data and the behaviour of the failure rate function, among others. The estimation of parameters is carried out by the method of maximum likelihood and a Monte Carlo simulation study is provided to assess the performance of this method. The practical usefulness of the model is illustrated by two real data sets, showing that it provides a better fit than other previously considered probability distributions.

CO0193: **Testing for the generalized Poisson-inverse Gaussian distribution**
Presenter: **Virtudes Alba-Fernandez**, University of Jaen, Spain
Co-authors: Maria Dolores Jimenez-Gamero, Apostolos Batsidis

The generalized Poisson inverse Gaussian (GPIG) family is a flexible family of distributions, useful for modelling count data with different tail heaviness. It includes the Poisson, Poisson-inverse Gaussian and discrete stable distributions as special cases. A new goodness-of-fit test is proposed for the GPIP family which is based on the following: since the probability generating function (PGF) of the GPIP family is the unique PGF satisfying certain differential equation, and the empirical PGF, as well as its derivatives, consistently estimate the PGF, and its derivatives, the empirical PGF should approximately satisfy such an equation. The proposed test statistic is based on sizing the coefficients of the polynomials up in the cited empirical version of the equation. It is shown that the test is consistent against fixed alternatives. The null distribution of the test statistic can be consistently approximated by means of a parametric bootstrap and a weighted bootstrap. The finite sample performance of the proposed test is investigated by means of a simulation study, where the goodness of the proposed approximations is numerically studied and the test is compared, in terms of power, to others.

CO0203: **Bayesian estimation techniques in certain life-testing models**
Presenter: **Inmaculada Barranco-Chamorro**, University of Sevilla, Spain

In life-testing and reliability models, mainly due to economic reasons, the experiments must finish before all the units in the sample fail. So it is common to deal with censored data. One of the most important censoring schemes, proposed in recent years, is progressive type-II right censoring in which surviving units can be censored at any time of failure. Under this censoring scheme, bayesian results are given when sampling from the Burr type-XII and the generalized half-logistic distribution. Both models are widely used in life-testing. Expressions are proposed for the Bayes and posterior risks of Bayes estimators under different loss functions. A simulation study is carried out to show the importance of the different features involved in the process of estimation.

CO0332: **Bayesian estimation of independence and compositional interaction in a two-way classification**
Presenter: **Mi Ortego**, Universitat Politecnica de Catalunya, Spain
Co-authors: Juan Jose Egozcue

A two-way discrete classification is characterized by a table of probabilities. This table of probabilities can be assumed to be the parameters of a multinomial sampling. This two-way probability table can be interpreted as a composition in the simplex, and therefore the Aitchison geometry is a suitable structure for its analysis. In this geometry, the nearest independent table to a probability table is the one built from the geometric marginals. The compositional difference between the original and the independent table is called interaction table, which encloses the information about the dependence between the two classifications. The independent and the interaction table are an orthogonal decomposition of the table of probabilities and can be represented using the centered-logratio transformations of both tables. The square Aitchison norm of the interaction table is called simplicial deviance and it is a dependence measure. Starting from a contingency table under a multinomial sampling, a Bayesian procedure is proposed in order to obtain the orthogonal decomposition into its independent and interaction tables. Interaction tables can be simplified in order to improve their interpretation. A Bayesian assessment of independence is used to this end.

ADVANCES IN ROBUST STATISTICS
CO021 **Room Sala 2** **Chair: Marco Riani**

CO0300: **Finding the number of groups in model-based clustering via constrained likelihoods**
Presenter: **Luis Angel Garcia-Escudero**, Universidad de Valladolid, Spain
Co-authors: Andrea Cerioli, Agustin Mayo-Iscar, Marco Riani

One of the most difficult problems in clustering is how to choose the number of clusters k. In model-based clustering, it is quite common the use of complexity-penalized likelihoods where the penalty term takes into account the number of free parameters. For instance, the BIC and ICL criteria can be used depending on whether mixture or classification likelihoods are considered. Unfortunately, these likelihoods are unbounded. This can be solved by considering appropriate constraints on the clusters' scatter matrices which it also avoids traditional algorithms from being trapped in (spurious) local maxima. Controlling the maximal ratio between the eigenvalues of the scatter matrices to be smaller than $c(\geq 1)$ has been proposed. Developing the associated penalized likelihood criteria requires taking into account the higher model complexity that a higher c entails. Clustering should not be seen as a fully automatic task and any user has to play an active role by specifying somehow the desired type of partitions. This specification can be done by fixing c depending on the clustering application. A fully automatized procedure, leading to a small and ranked list of optimal (k,c), will be presented. Extension to robust clustering will also be outlined.

CO0324: **The effect of trimming on algorithms for combining groups of pre-classified observations**
Presenter: **Andrea Cerasa**, Joint Research Centre, Italy
Co-authors: Andrea Cerioli

Three algorithms for merging homogeneous groups of pre-classified observations have been recently proposed. Their application on international trade data allows a synthetic representation of the market and a clear identification of anomalous commercial behaviors. To assess the performance of the algorithms we conducted Monte Carlo experiments. At that stage, methods for mitigating the effect of the possible presence of outliers had been left for forthcoming development of the procedure. Since anomalous declarations are quite usual in international trade data, non-robust procedures may result distorted and unfair trade strategies may be masked and remain undetected. We extend the simulation experiments by

introducing three different data contamination structures, in order to quantify the effects of the presence of anomalous observations on operationally relevant results. Moreover, we propose to pre-filter the data using a robust regression approach based on LMS and the Forward Search. The simulation results on filtered data help us measuring the effect of trimming on algorithms accuracy and choosing the best cleaning strategy for international trade data. An application to real data concludes the study.

CO0327: **Robust methods for analysis of 3-way compositional data in R**
Presenter: **Valentin Todorov**, UNIDO, Austria
Co-authors: Maria Anna Di Palma, Michele Gallo

The standard multivariate analysis addresses data sets represented as two dimensional matrices. In recent years, an increasing number of application areas like chemometrics, computer vision, econometrics and social network analysis involve analysis of data sets that are represented as multidimensional arrays and multiway data analysis becomes popular as an exploratory analysis tool. The most popular multiway models are CANDECOMP/PARAFAC and TUCKER3. The standard algorithms for computing these models are based on alternating least squares (ALS) and thus are vulnerable to the presence of outlying data points. Even a single outliying data point can strongly influence the resulting model and the conclusions based on it. Therefore robust methods are preferred. Additional difficulties for the analysis present cases of compositional data which consist of vectors of positive values summing to a unit, or in general, to some fixed constant for all vectors. They appear as proportions, percentages, concentrations, absolute and relative frequencies. We present a robust version of Tucker3 which is extended to handle compositional data. This method, together with a robust version of PARAFAC, also with an option for handling compositional data are implemented in an R package for analysis of multiway data sets.

CO0223: **Robustness for multilevel models with the forward search**
Presenter: **Luigi Grossi**, University of Verona, Italy
Co-authors: Aldo Corbellini, Fabrizio Laurini

Robustness of standard regression models have been studied quite extensively. When repeated measures are available, the methodological framework is generalized to multilevel models, for which little is known in term of robustness, even in the simplest case of ANOVA. We present a sequential forward search algorithm for multilevel models that allows robust and efficient parameters estimation in presence of outliers, and it avoids masking and swamping. The influence of outliers, if any is inside the dataset, will be monitored at each step of the sequential procedure, which is the key element of the forward search. There are peculiar features when the forward search is applied to multilevel models. Such features pose new computational challenges, as some restrictions, that make the sub-models identifiable at every step. Preliminary results on simulated data have highlighted the benefit of adopting the forward search algorithm, which can reveal masked outliers, influential observations and show hidden structures. An application to real data is also illustrated, where trades of coffee to European countries are analyzed to identify outliers that might be linked to potential frauds.

NONPARAMETRIC METHODS

CG011 **Room Sala 3** **Chair: Valentin Zelenyuk**

CC0379: **A test based on kernel density estimation for the eigenvalues in two-sample problem**
Presenter: **Hidetoshi Murakami**, Tokyo University of Science, Japan

A hypothesis test is considered for the eigenvalue of covariance matrix in two-sample. Though the test statistic can be constructed by a parametric procedure, large samples are necessary to keep a significance level. Since it is difficult to derive the exact distribution of the eigenvalues of the covariance matrix, we can consider a nonparametric procedure. We propose a statistic based on kernel density estimation to test the hypothesis that the jth largest eigenvalues of a covariance matrix are equal in two-sample. By simulations, we investigate the actual significance level and the power of the procedure using several tests for variance, and find that the proposed test is suitable for various cases.

CC0374: **Nonparametric estimation of ROC surfaces under verification bias**
Presenter: **Khanh To Duc**, University of Padua, Italy
Co-authors: Monica Chiogna, Gianfranco Adimari

In three-class diagnostic problems, ROC surfaces are commonly used for the evaluation of diagnostic markers. When all subjects are verified, the ROC surface could be constructed by nonparametric estimates of true class fractions. However, sometimes it is not feasible to obtain disease status verification for all study subjects, due to the expensiveness or invasiveness of the gold standard test. In such situations, the estimates based only on the verified subjects are typically biased. This bias is known as verification bias. In the last fifteen years, various methods have been developed to deal with the verification bias problem, most of which assume that the true disease status, if missing, is missing at random. However, the majority of previous work treated the issue of correcting for the verification bias in ROC curves, whereas ROC surface analysis is very scarcely considered in the statistical literature. We discuss how to construct the ROC surfaces of continuous-scale diagnostic tests in the presence of verification bias. Our approach is based on nearest-neighbor imputation and adopts generic smooth regression models for both the disease and the verification processes. Consistency and asymptotic normality of the proposed estimators are proved. An illustrative example is also presented.

CC0416: **Smoothing parameters for recursive kernel density estimators under double truncation**
Presenter: **Yousri Slaoui**, University of Poitiers, France

A data-driven bandwidth selection procedure is proposed for the recursive kernel density estimators under double truncation. We show that, using the selected bandwidth and a special stepsize, the proposed recursive estimators will be quite better than the nonrecursive in terms of estimation error and much better in terms of computational costs. We corroborate these theoretical results through a simulation study.

CC0377: **Nonparametric dynamic discrete choice models for time series data**
Presenter: **Valentin Zelenyuk**, University of Queensland, Australia
Co-authors: Byeong Park, Leopold Simar

The non-parametric quasi-likelihood method is generalized to the context of discrete choice models for time series data and, in particular, when lags of the discrete dependent variable appear among regressors. We derive consistency and asymptotic normality of the estimator for such models for general case and illustrate it with a few simulated and real data examples.

SAMPLING AND SMALL AREA ESTIMATION

CG094 **Room Sala 4** **Chair: Isabel Molina**

CC0483: **Numerical optimization for multivariate optimal allocation problems with several levels of strata**
Presenter: **Martin Rupp**, University of Trier, Germany
Co-authors: Ralf Muennich

The aim of modern surveys is to provide accurate information on a large variety of variables as well as on different regional levels and other subclasses of the population. Hence, optimizing a stratified sampling design, optimal allocation of the sample size has to consider a vast number of strata along with optimization conflicts due to the complementary information of the variables of interest with regard to different levels of strata. Furthermore, particular quality or cost restrictions might be taken into account. Modelling this multivariate optimal allocation problem leads to a high dimensional multi-objective optimization problem with equality constraints and box-constraints for the stratum specific sample sizes.

Taking advantage of the special structure of the variance functions and applying Pareto-optimization, the problem can be equivalently reformulated as a significantly lower dimensional non-linear system of equations, depending only on the Lagrange multipliers. Even though this system is non-smooth, it can be solved applying a semi-smooth newton algorithm with appropriate starting point and step size strategies. Due to the lower dimension, computational time is reduced considerably. The performance of the developed algorithm is tested on a business data set.

CC0490: **Integer-valued algorithms for constrained optimal allocations in stratified sampling**
Presenter: **Ulf Friedrich**, Trier University, Germany
Co-authors: Ralf Muennich, Sven de Vries, Matthias Wagner

In stratified random sampling, minimizing the variance of a total estimate leads to the optimal allocation by Neyman and Tschuprow. This original method is in practice rarely appropriate since in many applications constraints on the sizes of certain strata have to be considered. Moreover, classical algorithms for this allocation problem yield real-valued rather than integer-valued solutions. When a rounding strategy is applied to obtain an integral solution, the rounded solution is not optimal in general and often infeasible. The integral allocation problem with upper and lower bounds is modeled as a separable and convex optimization problem and new algorithms are presented for its solution. The methods exploit the special polyhedral structure of the set of feasible allocations and share the important feature of computing the globally optimal, integral solution. It is proved that the problem is solvable in polynomial time complexity using these methods. Finally, the practical relevance is illustrated by solving numerical examples with several thousand strata and many constraints.

CC0484: **Small area estimation of biodiversity measures**
Presenter: **Philip Rosenthal**, Trier University, Germany
Co-authors: Jan Pablo Burgard, Stephan Feldmeier, Ralf Muennich, Michael Veith

Sustainability management and the protection of biodiversity are of increasing importance for evidence based policy. As policy can only be implemented in regions in reach of the policy maker good regional estimates for biodiversity measures are needed. Hill numbers provide a framework for such biodiversity measures. Because sampling of species is very costly and often restricted to certain regions, no traditional sampling is performed for many species. The data gathering process is usually not describable in the classical sampling theoretical framework. Observations are often sparse and scattered on the whole region of interest, leading to inaccurate outcomes of regional biodiversity measures when using classical estimation techniques. Model based small area methods may lead to more reliable estimates. We extend a multinomial logit model by a zero-inflated Poisson part. This model is used to construct a synthetic estimator for the Hill numbers for an arbitrary aggregation level of the observation grid. A parametric bootstrap is used for the mean squared error estimation. To validate the method a model based simulation is used to show the performances of the new estimator in different extreme scenarios. The new method is then applied to a grasshopper dataset to estimate their biodiversity on natural regions.

CC0481: **Non-linear small area models under constraints**
Presenter: **Julian Wagner**, University of Trier, Germany
Co-authors: Ralf Muennich

Small area estimation methods are applied to obtain reliable estimates for parameters of sub-populations with sub-sample sizes too small for direct estimates. Traditional small area models are based on the assumption of an approximately linear relationship within the data, which is sufficiently often violated in practice. A non-linear small area model based on spline approximation is presented, which allows for additional shape constraints on the regression function. Moreover, constraints on the small area estimates are taken into consideration leading to quadratic programming problems. The applicability of the method is shown by a practical application and the results are compared to different small area methods. Finally, an estimator of the small area prediction mean squared error is proposed and the extension of the model to a multidimensional setting is discussed.

BOOTSTRAP IN TIME SERIES ANALYSIS

CG026 **Room Sala 7** **Chair: Roland Fried**

CC0527: **Detection performance of likelihood ratio test for change-points based on bootstrap for AR(1) models**
Presenter: **Ceyda Yazici**, Middle East Technical University, Turkey
Co-authors: Ceylan Yozgatligil, Inci Batmaz

The detection of change-points in time series is an important issue especially in economics, finance, meteorology and energy. Change in mean, change in variance or any sudden increase or decrease in the series can cause breakpoints. In AR(1) models, the likelihood ratio test is conducted to test for a single breakpoint. However, if the sample size is small or the location of the breakpoint is close to the end or the beginning of the series, the detection performance becomes worse. In order to increase the correct detection percentage of the likelihood ratio test in these cases, a bootstrap method for dependent data is applied and its performance is investigated when the change is only in the mean under several breakpoint scenarios. The test is applied to simulated data and the results are compared with the results obtained from tests in the literature.

CC0393: **Test of mean difference in longitudinal data based on block resampling approaches**
Presenter: **Hirohito Sakurai**, National Center for University Entrance Examinations, Japan
Co-authors: Masaaki Taguri

The focus is on a two-sample problem, and propose two block resampling testing methods with permutation analogy for comparing the difference of two means in longitudinal data when the data of two groups are not paired. In order to detect mean difference of two samples, we consider the following four types of test statistics: (i) sum of absolute values of difference between two mean sequences, (ii) sum of squares of difference between two mean sequences, (iii) estimator of area-difference between two mean curves, and (iv) difference of kernel estimators based on two mean sequences. The considered block resampling techniques include circular block bootstrap and stationary bootstrap, and are used to approximate the null distributions of the above test statistics. Monte Carlo simulations are conducted to examine the size and power of the testing methods.

CC0354: **Abrupt change in mean avoiding variance estimation without boundary value problem and block bootstrap**
Presenter: **Barbora Pestova**, The Czech Academy of Sciences, Czech Republic

Sequences of weakly dependent observations that are naturally ordered in time are considered. Their constant mean is possibly subject to change at most once at some unknown time point. The aim is to test whether such an unknown change has occurred or not. The change point methods presented here rely on ratio type statistics based on maxima of cumulative sums. These detection procedures for the abrupt change in mean are also robustified by considering a general score function. The main advantage of the proposed approach is that the variance of the observations neither has to be known nor estimated. The asymptotic distribution of the test statistic under the no change null hypothesis is derived and is free of any tuning parameters. Moreover, we prove the consistency of the test under the alternative. A block bootstrap method is developed in order to improve computational performance of the asymptotic methods. The validity of the bootstrap algorithm is shown. The results are illustrated through a simulation study.

CC0560: **The comparison of block bootstrap techniques in case of generalized resampling scheme**
Presenter: **Lukasz Lenart**, Cracow University of Economics, Poland

Generalized Resampling Scheme (GRS) was recently introduced for nonstationary time series. In special case this procedure reduces to usual subsampling technique. Generally, GRS can be reducing to some new resampling methods. The consistency of GRS holds under general assumptions

concerning moment and mixing conditions. The aim is to compare in testing problem a few block bootstrap methods defining on a basis of GRS. This testing problem concerns frequency identification in mean function in class of nonstationary Almost Periodically Correlated time series. Usual MC simulations are carried out in order to examine the sizes and powers of the test.

Wednesday 24.08.2016 11:00 - 12:50 Parallel Session G – COMPSTAT

RECENT ADVANCES IN MIXTURE MODELING

CO053 **Room Sala 1** **Chair: Xinyuan Song**

CO0170: **Selection of the number of components in mixture regression model**
Presenter: **Wing Kam Fung**, University of Hong Kong, China

Mixture regression model has been proven to be a useful tool in the study of heterogeneous populations arising in many fields. An important question in model building is the selection of the number of components. The majority of existing methods emphasize the goodness of fit and do not differentiate this problem with the diagnosis of other aspects of a mixture model. The classification probability instability method based on the concept of instability is introduced to select the number of components in mixture regression models. The proposed method can deal with the situation where the number of components is only one, in which the existing instability procedures may not be able to investigate. Stages are used to handle the possible multilevel structure in the components. Two variations are examined for small samples. The higher accuracy compared with some information criterion procedures are illustrated by simulations. A real data set on plasma beta-carotene concentration was analyzed. The selected mixture model offers a better interpretation about the relationship between the plasma beta-carotene concentration and dietary factors and personal characteristics.

CO0204: **Scalable clustering methods for dynamic Poisson graphical models**
Presenter: **Yingying Wei**, The Chinese University of Hong Kong, China
Co-authors: Xiangyu Luo

Graphical models describe the dependence structures among random variables. Recently there has been active research on modeling multiple Gaussian graphical models together. Nevertheless, there is a lack of research on Poisson Graphical models which describe counts data, let alone jointly modeling multiple Poisson Graphical models. We present a novel Bayesian nonparametric dynamic Poisson graphical model for multivariate counts data. The model was motivated by the transcription factors (TF) networks that control gene expression. The TF networks are dynamically varying across diverse biological conditions and heterogeneous across the genome within each given condition. Our proposed model automatically captures the between-condition dynamics and within-condition heterogeneity. Despite the motivating example, the proposed model is applicable to a large class of multivariate counts data. A parallel Markov Chain Monte Carlo algorithm is developed for posterior computation, which enables the computation to be scaled up efficiently.

CO0185: **A pivotal allocation based algorithm for solving the label switching problem in Bayesian mixture models**
Presenter: **Han Li**, Shenzhen University, China

In Bayesian analysis of mixture models, the label switching problem occurs as a result of the posterior distribution being invariant to any permutation of cluster indices under symmetric priors. To solve this problem, we propose a novel relabeling algorithm and its variants by investigating an approximate posterior distribution of the latent allocation variables instead of dealing with the component parameters directly. We demonstrate that our relabeling algorithm can be formulated in a rigorous framework based on information theory. Under some circumstances, it is shown to resemble the classical Kullback-Leibler relabeling algorithm and include the recently proposed Equivalence Classes Representatives relabeling algorithm as a special case. Using simulation studies and real data examples, we illustrate the efficiency of our algorithm in dealing with various label switching phenomena.

CO0398: **Using intraclass correlation coefficients to quantify spatial variability of catastrophe model errors**
Presenter: **Baldvin Einarsson**, AIR-Worldwide, United States
Co-authors: Rafal Wojcik, Jayanta Guin

Systematic spatial errors of natural catastrophe (CAT) models are quantified using hierarchical linear models. Insurance claims are grouped into spatial bins on a regular grid, which avoids computationally expensive distance calculations when estimating spatial covariances. For insurance claims and CAT model estimates, damage ratios are used to determine the model errors. The spatial structure of claims distributions around a model estimate is determined via intraclass correlation coefficient (ICC). A methodology is introduced to incorporate all claims, which greatly enhances the usability and robustness of the statistical models. These statistical models can have a hierarchy of spatial bins nested within larger bins, and both the number of such hierarchies, as well as the sizes of the rectangular bins at each layer, are investigated. Furthermore, several validation procedures are presented using the claims data from a major earthquake. The results are obtained with the R-package lme4.

CO0258: **A general non-linear multilevel structural equation mixture model**
Presenter: **Augustin Kelava**, Eberhard Karls Universitaet Tuebingen, Germany
Co-authors: Holger Brandt

In the past two decades latent variable modeling has become a standard tool in the social sciences. In the same time period, traditional linear structural equation models have been extended to include either a) nonlinear interaction and quadratic effects, b) multilevel effects, or c) mixtures. In recent years, (linear) parametric multilevel structural equation mixture model frameworks have been presented and made available in popular statistical latent variable modeling software. Nevertheless, these frameworks are restricted to parametric linear relationships. A general nonlinear multilevel structural equation mixture model (GNM-SEMM) is presented that combines recent semiparametric nonlinear structural equation models with multilevel structural equation mixture models for clustered and nonnormally distributed data. The proposed approach allows for semiparametric nonlinear relationships at the within and at the between levels. Examples from the educational science are presented to illustrate different submodels from the general framework.

ROBUST INFERENCE AND ROBUST STATISTICS WITH R

CO029 **Room Sala 2** **Chair: Valentin Todorov**

CO0229: **Estimating the number of clusters in OTRIMLE robust Gaussian mixture clustering**
Presenter: **Christian Hennig**, UCL, United Kingdom
Co-authors: Pietro Coretto

The method of Optimally Tuned Robust Improper Maximum Likelihood (OTRIMLE) has been recently developed for clustering data based on a Gaussian mixture model, but allowing for some observations that could not reasonably be assigned to any cluster. Those are modelled by a constant pseudo-density. This constant is found by minimising the Kolmogorow distance between the distribution of Mahalanobis distances to the corresponding cluster mean of the portion of the data classified as non-outlying and a chi squared distribution. We explore model diagnostic and estimation of the number of clusters by parametric bootstrap: we generate many datasets from the Gaussian mixture (non-outlying) part of the estimated mixture, using the estimated parameters, and we compare the distribution of the values of the Mahalanobis criterion to the value achieved for the dataset under study. This allows us to see which numbers of clusters yield models that are consistent with the data.

CO0266: **Sparse and robust PLS for regression and binary classification**
Presenter: **Peter Filzmoser**, Vienna University of Technology, Austria
Co-authors: Irene Hoffmann, Sven Serneels, Christophe Croux, Kurt Varmuza

Partial least squares (PLS) regression is successfully used to regress a univariate response on a potentially big number of explanatory variables. PLS can also be used in a high-dimensional two-group discrimination setting; in this case the response is a binary variable representing the two groups. The key idea is to reduce the dimensionality of the regressors by projection to latent structures. Since the dimension reduction and the regression step are sensitive to outlying observations or heavy-tailed distributions, a robust method called Partial Robust M-estimation (PRM) has been introduced which robustifies both steps. In high-dimensional regression or classification problems, variable selection is frequently desired, since it simplifies the interpretation of the resulting model and stabilizes the prediction model. For this reason, a sparse version of PLS has been introduced which leads to a regression coefficient vector that contains zeros. We propose a robust version, called sparse PRM, which turns out to be very useful in high-dimensional regression problems in presence of data artifacts. This method has been modified to work for binary classification problems as well. Inference based on bootstrap allows to identify the significant predictors.

CO0296: **Robust clustering for functional data**

Presenter: **Diego Rivera Garcia**, Centro de Investigacion en Matematicas, Mexico

Co-authors: Luis Angel Garcia-Escudero, Agustin Mayo-Iscar

Many algorithms for clustering analysis when the data are curves or functions have been proposed recently. However the presence of contamination in the data can influence the performance of most of these clustering techniques. Therefore, it would be interesting to get available tools for robustifying clustering algorithms. We propose a robust clustering method based on approximate coordinates obtained by applying functional principal components. This robustness is based on the joint application of trimming, for reducing the effect of contaminated observations, and constraints on the variances, for avoiding spurious clusters in the solution. The proposed method was evaluated through a simulation study, which showed an improved performance when compared with other recent methods for functional data clustering.

CO0338: **Compositional tables with applications in robust statistical analysis: Methodology and computing**

Presenter: **Kamila Facevicova**, Palacky University Olomouc, Czech Republic

Co-authors: Karel Hron, Valentin Todorov, Matthias Templ

Compositional tables represent a continuous counterpart to the contingency tables. Accordingly, their cells, containing in general positive real numbers rather than just counts, carry relative information about relationships between two factors. Consequently, compositional tables can be considered as a generalization of (vector) compositional data. Due to relative character of these observations, compositions are popularly expressed in orthonormal coordinates using sequential binary partition (SBP) prior to further processing using standard statistical tools. The contribution presents a general system of orthonormal coordinates with respect to the Aitchison geometry of compositional data, which is constructed as a combination SBPs of whole rows and columns of the table and which enables to analyze interactions between factors in a compositional table. The interpretation of coordinates is closely connected to odds ratios, which are popular also in context of contingency tables. The aim is to apply robust exploratory analysis like outlier detection and PCA including the respective visualization tools to a sample of compositional tables, with a particular focus on proper choice of interpretable orthonormal coordinates and assumptions for the corresponding robust estimators. Computations are performed using comprehensive functions from newly developed R-package on compositional tables.

CO0370: **TCLUST extensions**

Presenter: **Agustin Mayo-Iscar**, Universidad de Valladolid, Spain

Co-authors: Luis Angel Garcia-Escudero, Alfonso Gordaliza, Francesca Greselin

TCLUST is a model-based clustering robust methodology for multivariate data. Its robustness is based in the joint application of trimming and constraints. TCLUST extensions have been developed for estimating mixtures of linear models, mixtures of factor analyzers and mixtures of Skew-Normal models. In every of these frameworks, this methodology shows good performance when applied to data sets containing contaminated observations. Data driven tools for helping to the users in choosing the input parameters, related with the joint application of these tools, are also available.

MISSING DATA AND IMPUTATION

CO023 **Room Sala 4** **Chair: Faisal Maqbool Zahid**

CC0253: **Test of missing data mechanisms: An alternative to the Little test based on regression**

Presenter: **Serguei Rouzinov**, University of Lausanne, Switzerland

Co-authors: Andre Berchtold

Missing Data (MD) are common and occur in almost all surveys. The first step when working with MD is to determine whether they can be considered as missing completely at random (MCAR), missing at random (MAR), missing not at random (MNAR), or even a mixture of them (mixed MD mechanism). Nowadays, one of the most used methods for testing the MD mechanism is the Little likelihood ratio test. The aim of this method is to test whether a multivariate dataset has MCAR MD or not. However, this test has several limits. On the one hand, it is appropriate only for numerical data. On the other hand, it loses accuracy in presence of a mixed MD mechanism. An alternative approach based on a multiple linear regression model is proposed, whose dependent variable contains a certain amount of MD. The test draws conclusions from the comparison between the values predicted for the missing and the non-missing parts of the dependent variable. This test can be applied on categorical and numerical variables and it shows interesting properties in presence of a mixed MD mechanism. The test is described in details and numerical simulation results are provided.

CC0552: **Multiple imputation using sequential regression for high-dimensional data**

Presenter: **Faisal Maqbool Zahid**, Ludwig-Maximilians-University Munich Germany, Germany

Co-authors: Christian Heumann

Missing data is a ubiquitous in almost every field of research. Multiple Imputation (MI) is a commonly used technique to fill missing data with plausible values. It is common in applied research to face large number of variables for moderate number of cases. In such situations, the existing standard MI approach either performs poorly or fails to respond for $p > n$. A very limited literature is available to cope the issue while using MI. The question still of interest is about the best possible strategy to impute the missing data multiply with high-dimensional data. To address the issue, we are proposing a penalized version of standard MI. We are using lasso regression with high-dimensional data to impute the missing values. We compare the performance of our algorithm, with increasing dimension of the data, with some existing algorithms in different simulation studies. The performance is compared using Mean Squared Imputation Error (MSIE) and Mean Absolute Imputation Error (MAIE). The results of the study suggest that using lasso based MI approach is better option for imputation with high-dimensional data.

CC0361: **Variable selection for longitudinal biomarkers constrained by a detection limit**

Presenter: **Julia Geronimi**, CNAM, France

Co-authors: Gilbert Saporta

Repeated measures over time are common in the biomedical field, and widely used to analyze the link between covariates and a clinical criterion. In a longitudinal context, a high number of variables associated with the presence of missing data, are complex issues to be resolved. We deal with several types of covariates, some suffer from haphazard missingness, and others are subject to detection thresholds. For the latter, Tobit regression combined with bootstrap is an unbiased approach, but it needs complete predictors for the mean model. An adaptation of the well-known multivariate imputation by chained equation is proposed. We use the Tobit model as the imputation method for covariates below the

detection limit, predictive mean matching and logistic regression for others. Variable selection is done by using MI-PGEE which consists in the following ingredients: a) a group LASSO penalty is imposed on the group of estimated regression coefficients of the same variable across multiply-imputed datasets leading to a consistent selection. The optimal shrinkage parameter is chosen by minimizing a BIC-like criterion. b) GEE allows integrating correlations due to the longitudinal context. The usefulness of the new method is illustrated by an application on the FNIH project of the Osteoarthritis Initiative.

CC0320: **Single imputation by data depth**
Presenter: **Pavlo Mozharovskyi**, Centre Henri Lebesgue, France
Co-authors: Julie Josse, Francois Husson

Single imputation is an appropriate technique to handle missing data if one simply needs to complete a single data set, when no inference is required, when the applied statistical method is computationally too demanding for multiple data sets, or when a few values are missing only but one seeks an alternative to the list-wise deletion. The presented methodology for single imputation of missing values borrows the idea from data depth - a measure of centrality defined for an arbitrary point of the space with respect to a probability distribution or a 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. Being able to grasp the underlying data topology, the procedure is distribution free, allows to impute close to the data, preserves prediction possibilities different to local methods (nearest neighbor imputation, random forest), and has attractive robustness and asymptotic properties under elliptical symmetry. It is shown that its particular case - when using Mahalanobis depth - has direct connection to well known treatments for multivariate normal model, such as iterated regression or regularized PCA. Simulation and real data studies contrast the procedure with existing popular alternatives.

CC0551: **Inference when using nearest neighbors methods and the bootstrap**
Presenter: **Shahla Ramzan**, Ludwig Maximilians University Munich, Germany
Co-authors: Christian Heumann, Gerhard Tutz

Imputation is an attractive approach for filling the missing data values with their estimates. A number of methods are available in literature that can be used for imputing the missing data. However it is not advisable to treat the imputed data just as the complete data. Applying the existing methods to analyse the imputed data, for example, to estimate the variance and/or statistical inference will probably produce invalid results because these methods do not account for the uncertainty of imputations. We present analytic techniques for inference from a dataset in which missing values have been replaced by nearest neighbors imputation method. A simple and easy to use bootstrap algorithm that combines the nearest neighbors imputation with bootstrap resampling estimation to obtain valid bootstrap inference in a linear regression model is suggested. More specifically, imputing the bootstrap samples in the exact same way as original data was imputed produces correct bootstrap estimates. Simulation results show the performance of our approach in different data structures.

NEW ADVANCES IN MULTISET AND MULTIWAY DATA ANALYSIS I

CO041 **Room Sala 3** **Chair: Katrijn Van Deun**

CO0278: **Overlapping clusterwise simultaneous component analysis**
Presenter: **Eva Ceulemans**, University of Leuven, Belgium
Co-authors: Kim De Roover, Paolo Giordani

When confronted with multivariate multiblock data (i.e. data in which the observations are nested within different data blocks that have the variables in common), it can be useful to synthesize the available information in terms of components and to inspect between-block similarities and differences in component structure. To this end, the clusterwise simultaneous component analysis (C-SCA) framework was developed across a series of papers: C-SCA partitions the data blocks into a limited number of mutually exclusive groups and performs separate SCAs per cluster. We present a more general version of C-SCA. The key difference with the existing C-SCA methods is that the new method does not impose that the clusters are mutually exclusive, but allows for overlapping clusters. Therefore, the new method is called Overlapping Clusterwise Simultaneous Component Analysis (OC-SCA). Each of these clusters corresponds to a single component, such that all the data blocks that are assigned to a particular cluster have the associated component in common. Moreover, the more clusters a specific data block belongs to, the more complex the underlying component structure. A simulation study and an empirical application to emotion data are included.

CO0286: **Partitionned matrices in multiset and multiway data analysis**
Presenter: **Pasquale Dolce**, ONIRIS-Nantes, France
Co-authors: Mohamed Hanafi, David Legland

It is well known that matrix algebra contributes to consolidate the mathematical basis of the multiset and multiway data methods, as well as to provide computational and geometric tools for the implementation of dedicated software. However, matrices and tensors do not cover the full range of data acquired in multiset and multiway data analysis. Indeed, the continuous development of sensors allows to acquire various kinds of measurement on one or more groups of individuals. Mathematically, these data can be conceptualized as partitioned matrices. Both matrix algebra and tensor algebra appear inadequate to manipulate partitioned matrices. The manipulation of partitioned matrices as simple matrices, without considering the associated partition, is inappropriate when the partition is an essential component of the problem to be solved, as for the so-called multiblock methods. Moreover, the partitioned matrices cannot be represented as tensors since the modes of blocks are all not identical. As a result, new specific rules for computation on partitioned matrices are needed. An appropriate vocabulary is introduced and discussed for partitioned matrices, standard notations for these entities and, in particular, different types of products between partitioned matrices. Furthermore, a computational kernel for handling partitioned matrices under R and Matlab environment will be presented.

CO0295: **Mixture simultaneous factor analysis and Wald tests for factor loading differences in multivariate multilevel data**
Presenter: **Kim De Roover**, KU Leuven, Belgium
Co-authors: Jeroen Vermunt, Marieke Timmerman, Eva Ceulemans

Multivariate multilevel data consist of multiple data blocks involving the same variables; for instance, inhabitants from different countries rating emotion norms. The associated research questions often pertain to the underlying covariance structure (e.g. which dimensions underlie the individual scores), and whether it holds for each data block (e.g. do the underlying dimensions differ across countries). To answer such questions, mixture simultaneous factor analysis (MSFA) performs a mixture clustering of the blocks according to their factor structure, ignoring mean differences. Specifically, MSFA assumes that the data are sampled from a mixture of multivariate normal distributions with different covariance matrices, modeled by a low rank factor model, and that all individuals within a block are sampled from the same distribution. Comparing the cluster-specific factor loadings, one may either detect that the data blocks differ completely in underlying dimensions, or that only a few variables behave differently in some blocks. By simulations and empirical analysis using Latent Gold, we evaluate how well Wald-tests for loading differences can determine which variables make out the between-block structural differences.

CO0322: **Regularised multiblock methods for cancer patient classification**
Presenter: **Tommy Lofstedt**, Umea University, Sweden
Co-authors: Patrik Brynolfsson, Thomas Asklund, Tufve Nyholm

Regularised generalised canonical correlation analysis (RGCCA) was recently extended to allow for the connections between blocks to go beyond

the covariance link, and to predict a single binary outcome variable through logistic regression. The generalisation is called multiblock logistic regression (Multiblog) and contains logit links between the blocks and the outcome, and covariance links between the blocks. We extend the Multiblog model by allowing the path coefficients to be $c_{ij} \geq 0$, to control the influence of the covariance on the logistic regression models, and by regularising the regression coefficients. The aim was to study gray-level co-occurrence matrices (GLCMs) computed from three types of MRI images from 23 patients with high-grade glioma. The MRI images were the estimate of the apparent diffusion coefficient (ADC), T1- and T2-weighted images. These three types constituted three blocks, and the purpose was to predict whether the tumour would regress or progress, or whether the patient would exceed the median survival time or not. We compared the performance of these new extended RGCCA models and of several classical machine learning and multiblock methods, with or without regularisation, when applied on the GLCM data to those of the traditional Haralick feature extraction approach.

CO0328: **A general overview of multiblock component methods**
Presenter: **Arthur Tenenhaus**, Laboratoire Signaux et Systemes, France
Co-authors: Michel Tenenhaus, Patrick Groenen

A new framework for multiblock component methods is proposed. In that framework, two cases are considered: blocks are either fully connected or connected to the superblock (concatenation of all blocks). This framework relies on a new version of regularized generalized canonical correlation analysis (RGCCA) where various scheme functions and shrinkage constants are considered. The proposed iterative algorithm is monotone convergent and guarantees obtaining a solution of the stationary equations of RGCCA. For the scheme functions and shrinkage constants equal to 0 or 1, many multiblock component methods are recovered.

TUTORIAL 2

CO102 **Room Sala Camara** **Chair: Maria Brigida Ferraro**

CO0582: **A methodology to analyze fuzzy data**
Presenter: **Maria Angeles Gil**, Universidad de Oviedo, Spain

Fuzzy data are often used to model data associated with intrinsically imprecise-valued magnitudes/attributes (say perceived quality, satisfaction, attitude, and so on) in a random environment. The aim is to recall the required preliminary tools, and to present some of the already established statistical developments in connection with the central tendency/location and dispersion/scale of random fuzzy numbers. The estimation and testing methods about the population Aumann-type mean(s) are to be exposed. Some alternate robust location measures are introduced, their estimation is examined and their robustness is discussed. Regarding dispersion, the estimation and testing methods about the population Frechet-type variance(s) are to be described and some alternate robust scale measures are introduced, their estimation is examined and their robustness is discussed. Some of the presented methods will be illustrated by means of a real-life example. This example will serve to show the convenience of using the scale of fuzzy numbers instead of other scales like Likert-type ones or their fuzzy linguistic counterpart in dealing with data from these intrinsically imprecise-valued magnitudes/attributes. Related studies as well as some future directions will be shortly commented.

MACHINE LEARNING

CC067 **Room Sala 5** **Chair: Stefan Wager**

CC0433: **On the hyperparameter settings of random forests**
Presenter: **Philipp Probst**, LMU Munich, Germany
Co-authors: Anne-Laure Boulesteix, Bernd Bischl

Due to their good predictive performance, simple applicability and flexibility, random forests are getting increasingly popular for building prediction rules. Unfortunately, not much knowledge is available about the ideal hyperparameter settings of random forests. Some important hyperparameters are the number of trees, the number of randomly drawn features at each split, the number of randomly drawn samples in each tree and the minimal number of samples in a node. Common modern strategies for tuning are grid search, random search, iterated F-racing or Bayesian optimization. This can be too complicated for users without expertise on random forests, computationally costly or even infeasible in case of too big datasets. In an empirical study, we study the influence of a diverse range of hyperparameter settings of random forest algorithms and implementations of many different R packages on more than 200 different regression and classification problems from the OpenML platform. We use out-of-bag predictions and different performance measures for evaluation, and simple meta-learning to relate the performance results to data set characteristics. Our results yield valuable insights into a) parameter sensitivity for different performance measures b) optimal default settings, to be applied without further tuning c) tuning starting points and ranges for less time-consuming model building.

CC0489: **Recovery of weak signal in high dimensional linear regression by data perturbation**
Presenter: **Yongli Zhang**, University of Oregon, United States

How to recover the weak signal (i.e. small nonzero regression coefficients) is a difficult task in high dimensional data with multicollinearity. Both exhaustive search and stepwise methods fail to select the true model as the nonzero coefficients are below some threshold. We propose a procedure, Perturbed Model Selection (PMS), to recover weak signal by adding random perturbations to the feature matrix. It is shown through theory and simulations that PMS achieves substantial improvement upon the chance of recovering informative features and outperforms other methods at a limited expense of computation. In theory, the aim is to derive a quantitative relationship between selection consistency and computing and to demonstrate the trade-off between them. The real data example revealed that PMS improved forecasting by combining the power of decorrelation and resampling.

CC0384: **Trees garrote for regression analysis**
Presenter: **Masatoshi Nakamura**, Oita university, United States

In regression analysis, stochastic models are often constructed to model relationships between outcomes and explanatory variables, and we derive statistical interpretations for data based on these models. However, if we use only linear regression models, constructing a true model reflecting actual characteristics can be difficult. A tree-structured approach is recommended, such as classification and regression trees (CART), which develops a tree and provides an interpretation of the data based on the fundamental model derived from the tree. Random Forest (RF) involves an ensemble learning method based on the trees and can predict outcomes more precisely. However, RF cannot provide a tree-structured model for interpreting the data. A nonnegative garrote (NNG), a shrinkage estimator, is examined, and trees garrote (TG) is proposed as an adjustment of RF based on NNG. Some shrinkage estimators for ensemble learning are reported to yield better predictive performance. In addition, TG can lead to tree-structured models that are useful for interpretation of data. Simulation studies show that the proposed method is highly accurate predictively. Finally, two case studies of diabetes and prostate cancer data illustrate descriptive features of tree-structured models based on TG.

CC0501: **Destination prediction by trajectory distribution-based models**
Presenter: **Brendan Guillouet**, Institut de Mathematiques de Toulouse, France
Co-authors: Jean-Michel Loubes, Philippe Besse, Francois Royer

A data-driven methodology to predict the final destination of vehicle trips based on their initial partial trajectories is introduced. First, a clustering of trajectories is produced using hierarchical clustering and based on a new distance between trajectories, the Symmetrized Segment-Path Distance, (SSPD). The clusters obtained describe the main patterns of the traffic flow based on the drivers' usage. Locations within each of these pattern are

then modeled by a mixture of $2d$ Gaussian distributions. Hence, a data driven grid is learned based on the behaviors of the drivers. Finally, this model is used to predict the final destination of a new trajectory based on their first locations with a two step procedures: the new trajectory are first assigned to the clusters it belongs the most likely. Secondly, the final destination is predicted using characteristics from trajectories inside these clusters. This methodology has been successfully tested on two different datasets, assessing its capacity to be adapted to different networks. One of these datasets comes from the ECML/PKDD 15: Taxi Trajectory Prediction Kaggle challenge. Our final results produce promising result taking into account that our model can be re-used directly for a different test dataset, and can also be used to predict the destination during trajectory completion, without requiring a new training.

CC0407: **A toolkit for stability assessment of tree-based learners**
Presenter: **Michel Philipp**, University of Zurich, Switzerland
Co-authors: Achim Zeileis, Carolin Strobl

Recursive partitioning techniques are established and frequently applied for exploring unknown structures in complex and possibly high-dimensional data sets. The methods can be used to detect interactions and nonlinear structures in a data-driven way by recursively splitting the predictor space to form homogeneous groups of observations. However, while the resulting trees are easy to interpret, they are also known to be potentially unstable. Altering the data slightly can change either the variables and/or the cutpoints selected for splitting. Moreover, the methods do not provide measures of confidence for the selected splits and therefore users cannot assess the uncertainty of a given fitted tree. We present a toolkit of descriptive measures and graphical illustrations based on resampling, that can be used to assess the stability of the variable and cutpoint selection in recursive partitioning. The summary measures and graphics available in the toolkit are illustrated using a real world data set and implemented in the R package **stablelearner**.

APPLIED STATISTICS

CG074 **Room Sala 7** **Chair: Angela Blanco-Fernandez**

CC0523: **Dynamic clustering of multiple multivariate time series: Application to climate data**
Presenter: **Ceylan Yozgatligil**, Middle East Technical University, Turkey
Co-authors: Sipan Aslan, Cem Iyigun

A new time series clustering approach which is a nonlinear time series model based dynamic clustering approach for multiple multivariate time series is proposed. Observing similarity-dissimilarity between time series is accomplished by evaluating the approximations to their unknown data generating mechanisms rather than comparing trace-like pattern similarities in the time series. In other words, the proposed approach is mainly aimed at forming and acquiring distinguishing information-features throughout the given period of a time series using the threshold autoregressive and threshold vector autoregressive models where the actual nature of the underlying process or data generating mechanism is not known. The effectiveness of the proposed approach is illustrated via simulation examples, and then, it is applied to the problem of defining climate regions based on multivariate meteorological time series.

CC0396: **Comparing matrix factorisation approaches to fuzzy clustering**
Presenter: **Abdul Suleman**, Instituto Universitario de Lisboa ISCTE-IUL BRU Lisboa Portugal, Portugal

An empirical study is presented to compare three algorithms for fuzzy clustering in the framework of nonnegative matrix factorisation: archetypal analysis (ARCH), factorised fuzzy c-means (F-FCM) and unconstrained least squares (ULSQ). As an initial step, we conduct a Monte Carlo simulation with artificial data which configure several cluster contexts according to membership degree, noise contamination and density. The goodness of fit of the estimated fuzzy partitions is assessed through a generalised version of the Dice index, given fuzzy class labels. The F-FCM performs better than others when the data have a clear cluster structure, i.e. high membership in clusters, regardless of the density pattern or amount of noise. In contrast, the other two algorithms outperform F-FCM when the data have a scattered distribution. The ARCH algorithm generally performs better than ULSQ and additionally provides more stable solutions. It is therefore preferable to use ARCH despite its higher computational effort. A second experiment is carried out using data arising from real life problems and devoted to classification task. We can further confirm the effectiveness of the F-FCM algorithm in dealing with this kind of data and thus recommend it for fuzzy clustering purposes.

CC0397: **NPC to assess effects of maternal iodine nutrition and thyroid status on children cognitive development**
Presenter: **Massimiliano Giacalone**, University of Naples - Federico II, Italy
Co-authors: Angela Alibrandi, Agata Zirilli, Maria Carla Moleti

Maternal iodine nutrition and thyroid status may inuence neurocognitive development in children. We investigate the effects on the intelligence quotient (IQ) of children born to mothers with different levels of iodine supplementation, with or without the administration of levothyroxine (LT4), prior to and during pregnancy. From a methodological point of view, we used the Non Parametric Combination test or NPC test, based on permutation solution. It was chosen for the several optimal properties of which it is characterized, that make it very flexible and widely applicable in many fields; in particular, it allows stratified analyses and represents an effective solution for problems concerning the testing of multidimensional hypotheses, that are difficult to face in a parametric context.

CC0400: **Detection of exceptional genomic words: A comparison between species**
Presenter: **Ana Helena Tavares**, University of Aveiro, Portugal
Co-authors: Vera Afreixo, Joao Rodrigues, Carlos Bastos, Armando Pinho, Paulo Ferreira, Paula Brito

The potentialities of the inter-word distances to detect exceptional genomic words (oligonucleotides) in several species is explored using whole-genome analysis. We confront the empirical results obtained from the complete genomes with the corresponding results obtained from the random background. We develop a procedure, based on some statistical properties of the global distance distributions in DNA sequences, to discriminate words with exceptional inter-word distance distribution and to identify distances with exceptional frequency of occurrence. We identify the statistically exceptional words in whole-genomes, i.e. words with unexpected inter-worddistance distributions, and we suggest species signatures based on exceptional word profiles.

CC0329: **The temperature of Indian cities: Some insights using change point analysis with functional data**
Presenter: **Poonam Rathi**, N/A, India

In recent years there has been considerable concern expressed worldwide regarding increase in temperature popularly called the global warming problem. However, not much work has been done on this in the Indian context. We examine monthly temperature data of five Indian cities for the period 1961 to 2013. We introduce a new change point detection method for functional data and use it to investigate the existence of change point for the temperature data series of five Indian cities namely Srinagar, Imphal, Trivandrum, Bengaluru and Ahmedabad. It is found that there has been a rise in the average temperature for all cities except Ahmedabad during this period. The magnitude of warming is found not to be uniform but vary across cities. The estimated change points for the four cities are not identical but all are in the period 1989 - 1998. The findings suggest that immediate policy measures are required to ensure that no further warming happens in these cities.

POSTER SESSION I

CP106 **Room Ground Hall** **Chair: Marta Garcia Barzana**

CP0431: **FinSam: An R package for finite population sampling analysis**
Presenter: **Jacinto Martin Jimenez**, Universidad de Extremadura, Spain
Co-authors: Manuel Molina, Juan Carlos Ridruejo Sayavera

FinSam is an R package intended to support estimation analysis in finite population sampling. It covers the usual basic sampling schemes like simple random sampling and unequal probabilities, trough Horvitz-Thompson estimator. Besides, it includes cluster sampling (one and two stages) and indirect methods, ratio and regression estimators. The package also includes some complementary material, like inverse sampling, randomized responses, Jackknife method and others. It contains data examples and functions that allow us to obtain the different estimators and their variances in the command-line frontend. The final objective of the package is to use it in an advanced course on Finite Population Sampling. So, it includes a user-friendly menu-driven interface. It allows us to read data from files or introduce them manually according to the selected sampling analysis. Each method has a sub-menu with several options, mainly: Type of sampling, Managing and visualizing data and Computations. This is the first version of the FinSam package, and we are willing to expand it with other methods and examples.

CP0559: **R package for statistical process control when data are functional: fda.qcr**
Presenter: **Salvador Naya**, University of A Coruna, Spain
Co-authors: Miguel Flores, Javier Tarrio-Saavedra, Ruben Fernandez Casal

Functional Data Analysis, Quality Control and Reliability (fda.qcr) package is a new R library that tackles the problem of statistical process control from the Functional Data Analysis approach. It implements different FDA techniques to perform exploratory analysis, iterative visualization of functional data, outlier detection using data depth approach, analysis of variance, linear model fitting with scalar or binary response, and statistical quality control using specific FDA control charts. To obtain the confidence bands of control charts for functional data, the functional quantiles (i.e. 0.975 and 0.025) corresponding to critical to quality (CTQ) functional variable are obtained. Method based on data depth approach, pointwise functional quantiles estimates, and smoothed bootstrap resampling are implemented. FDA control charts can be applied taking into account dependence between observations of a functional CTQ variable. The fda.qcr package implements block bootstrap method to estimate functional quantiles when observations are dependent.

CP0531: **A new R library for discriminate groups based on abundance profile and biodiversity on microbiome metagenomics matrices**
Presenter: **Toni Monleon-Getino**, Fundacion Bosch i Gimpera, Spain
Co-authors: Clara Rodriguez-Casado, Jorge Frias-Lopez

Metagenomics is the study of genetic material recovered directly from samples like microbiome which have shown an extraordinary diversity. While traditional microbiology and microbial genome sequencing rely upon cultivated clonal cultures, metagenomics gene sequencing cloned specific genes to produce a profile of diversity in the microbiome. There is a great interest in associating specific groups of organisms with health and disease typologies, but unfortunately there are very few statistical tools for profile analysis with the ability to differentiate types which can co-exist. We have developed MetagenOutlineLDA, a new library for R that allows a statistical analysis of metagenomic matrices using different statistical approaches. This library performs three basic tasks: 1) the estimation of metagenomic abundance profiles (relative abundance of species) for each sample using robust regression, 2) estimation of metagenomic biodiversity-alpha and 3) performing a discriminant analysis (LDA,SVM,NN, etc.) to distinguish groups (e.g. healthy /disease). Thus a new metagenomic analysis could indicate the group of belonging with a reasonable percentage. Examples of analysis using MetagenOutlineLDA with people affected by periodontitis and Crohn's disease are presented proving to be a useful library. The results confirm that the diseases studied not only alter the composition of the human microbiome, but also its structure.

CP0353: **An R commander plug-in for fitting generalized Waring regression models**
Presenter: **Maria Jose Olmo-Jimenez**, University of Jaen, Spain
Co-authors: Silverio Vilchez-Lopez, Antonio Jose Saez-Castillo

The generalized Waring regression model (GWRM) for overdispersed count data has the ability to split the variability of the response variable into three components: randomness, liability and proneness. This is one of the reasons why this model is being widely used in different fields. An R package, available in the comprehensive R archive network (CRAN), also called GWRM, was developed for fitting, describing and validating this model, but the use of the package requires to know the R programming language. In order to spread the use of the regression model and contribute to their knowledge among non-advanced users of R, a plug-in for R Commander, a basic statistics graphical user interface for R, is presented. The plugin, called RcmdrPlugin.GWRM, can be also downloaded from CRAN. To illustrate its usage and the menu options, an example is included.

CP0442: **An approach to the errors-in-variables regression model**
Presenter: **Taku Yamamoto**, Institute of Statistical Research, Japan
Co-authors: Makoto Muto, Teruo Nakatsuma

Various methods have been proposed in order to handle the errors-in-variables regression model where its explanatory variable contains a measurement error. A relatively mild condition will be assumed. Namely, the explanatory variable is assumed to be positively auto-correlated. This condition is satisfied in most economic time series. It is well known that the ordinary least squares estimator of the errors-in-variables model is asymptotically biased (inconsistent). The first purpose is to show that, when the (latent) explanatory variable is positively auto-correlated, the temporal aggregation of the model decreases the asymptotic bias and the mean squared error of the ordinary least squares estimator. However, the temporal aggregation cannot completely eliminate the asymptotic bias. The second purpose is to propose a convenient consistent estimator which suitably combines two biased estimators. It can be regarded as an alternative instrumental variable estimator. The suitably designed Monte Carlo experiments show the effectiveness of temporal aggregation in small samples, and that the proposed combined estimator is superior to the previously proposed consistent estimators. Finally, the proposed method is applied for testing the Fisher equation in Japanese economy.

CP0573: **LIHAR model for forecasting realized volatilities featuring long-memory and asymmetry**
Presenter: **JiWon Shin**, Ewha Womans University, Korea, South
Co-authors: DongWan Shin

It is well known the fact that financial time series have asymmetric variances and a recent paper revealed that an integrated HAR model beats the HAR model. So, we add a leverage term to an integrated HAR model, called as LIHAR model. Comparisons of forecasting ability of several models show that the IHAR model with leverage (LIHAR) is superior to the HAR and IHAR model. The model is applied for 20 real data sets of RV for financial indices DJIA, S&P 500, Russell 2000, KOSPI Composite, etc. The volatilities of the financial indices like stock price and foreign exchange are characterized by very persistent long memories and asymmetry. These features are so well-suited for the integrated heteroscedastic autoregressive model with leverage that the LIHAR model generally produces better out-of-sample forecasts than other models like HAR, IHAR and LHAR for the real data sets. This result supports that recent IHAR model is more suitable for forecasting RV for financial indices. It is good to consider long-memory, asymmetry, nonstationarity for forecasting realized volatilities.

CP0507: **Locally weighted mixture models for prediction from time series**
Presenter: **Najla Qarmalah**, Durham University, United Kingdom
Co-authors: Jochen Einbeck, Frank Coolen

Locally weighted mixture models for time series data are introduced and used to obtain predictions based on the fitted models. Given data of the form (t_i, y_i),$i = 1, ..., T$, we suppose a mixture model with the $k-$th component as $y_i = m_k(t_i) + \varepsilon_{ik}$ with mixing proportion $\pi_k(t_i)$, $k = 1, ..., K$ and K is the number of components. The $m_k(t_i)$ is a smooth unspecified regression function, and the error $\varepsilon_{ik} \sim N(0, \sigma^2)$ is independently distributed. Estimation of this model is achieved through a kernel-weighted version of the EM-algorithm, using exponential kernels with different bandwidths h_k which have bigger effect on the forecast. By modelling a mixture of local regressions at a target quantity t_T but with different bandwidths h_k, the estimated mixture probabilities are informative for the amount of information available in the data set at the scale of resolution corresponding to each bandwidth. Nadaraya-Watson and local linear estimators are used to carry out the localized estimation step. Based on the fitted model, several approaches for $m-$step-ahead predictions, $m = 1$ and 2, at time t_{T+m} are investigated. Real and simulated data are provided including data on energy use for Spain.

CP0496: **Dynamic binary choice panel data model with fixed effect based on transition model**
Presenter: **Takuma Kurosawa**, Tokyo University of Science, Japan
Co-authors: Asanao Shimokawa, Etsuo Miyaoka

In some medical research, we analyze dynamic binary choice panel data. Handling heterogeneity is one of the important things in considering these data. They are roughly divided into two types: a fixed effect and a random effect. We focus on dynamic binary choice panel data models with fixed effects. If only one lag observation which represent the dynamics is included the model, there are several proposed methods for analyzing the data. However, sometimes we are interested in more complex dynamics such as a model has more than one lag observations. To address this problem, we apply transition model to dynamic binary choice panel data and test efficiency of the model through some simulation studies. Moreover, we show the results of applying the model to actual data.

CP0399: **Various algorithmic approaches for the balancing problem**
Presenter: **Susie Fortier**, Statistics Canada, Canada
Co-authors: Michel Ferland

Time series data produced by National Statistical Offices and Systems of National Accounts must often respect a vast array of accounting relationships. These relationships can be quite simple, such as requiring that regional components add up to a national total or more complex, such as the econometric equalities that may be used to compute the Gross Domestic Product. As the data may come from various sources or undergo non-linear data processing such as seasonal adjustment, the accounting relationships must often be restored before publication. The process used to restore the accounting coherence in the data is referred to as balancing or reconciliation. The problem can be approached in various ways ranging from a purely numerical point of view to a fully parameterised model. Several solutions will be presented and discussed. From a regression-based model solved through matrix manipulation to a mathematical optimisation problem solved numerically, the algorithmic approaches will be compared with emphasis on their strengths and weaknesses. Statistics Canada's recent implementation of a numerical optimisation solution as part of their G-Series software will also be presented.

CP0157: **Asymptotic expansions for the estimators of Lagrange multipliers by the weighted score methods**
Presenter: **Haruhiko Ogasawara**, Otaru University of Commerce, Japan

Inverse expansions of parameter estimators are given in terms of their true values, where the estimators are obtained by the maximum likelihood and weighted score methods with constraints placed on the parameters using Lagrange multipliers. The corresponding expansions for estimated Lagrange multipliers are also given. These expansions are derived before and after studentization. The results with studentization give one-sided confidence intervals for the parameters up to third-order accuracy. As an application of the weighted score method, a modified Jeffreys prior to remove the asymptotic biases of the Lagrange multipliers as well as the parameter estimators is obtained under canonical parametrization in the exponential family.

CP0429: **Penalized splines with censored data**
Presenter: **Jesus Orbe**, University of the Basque Country, Spain
Co-authors: Jorge Virto

The problem of nonparametric curve fitting is considered in the specific context of censored data. That is, we do not observe completely the sample, because some of the data values are censored. This situation is very usual in survival or duration analysis. When the available data are complete, that is, without censored data, the problem of nonparametric curve fitting has been extensively studied and there is an enormous literature on this area. There are different methods, as for example, kernel smoothers and spline smoothers. Our proposal is situated under the splines approach. Thus, an extension of the penalized splines approach is provided for the case of censored data samples. This method works together the usual B-spline regression techniques and the smoothing spline approach reducing considerably the dimension of the estimation problem in large samples in comparison with the smoothing splines case. Using various simulation studies we analyze the effectiveness of the proposed method and show that the performance are quite satisfactory. Also a real data set is used to illustrate the proposed methodology and it is shown as a well alternative when we do not know the functional form on censored regression models.

CP0435: **A widely linear system for quaternion estimation**
Presenter: **Juan Carlos Ruiz-Molina**, University of Jaen, Spain
Co-authors: Jesus Navarro-Moreno , Rosa Maria Fernandez-Alcala , Jose Domingo Jimenez-Lopez

A Karhunen-Loeve series expansion is presented for quaternion representation and its potential application in the quaternion MMSE estimation problem is shown. The optimal implementation of the technique needs the computation of the true eigenvalues and eigenfunctions of an integral equation. This task can be very involved and sometimes even impossible. We avoid this shortcoming solving numerically the integral equation involved via the Rayleigh-Ritz method. Then the resulting approximate series expansion is used to propose a new form of the estimator. A numerical example illustrates the performance of the suggested solution.

CP0491: **A computational statistics approach for estimating the smallest natural number x for which $\pi(x) > li(x)$**
Presenter: **Ryuichi Sawae**, Okayama University of Science, Japan

In number theory, there may be an extremely large number, which is defined by the smallest natural number x such that $\pi(x) > li(x)$, where $\pi(x)$ is the prime counting function and $li(x)$ is the logarithmic integral function. The smallest natural number is known as the so-called Skewes' number, and has been proved to be between the 19th power of 10 and about 6.6587×10^{152} by many recent researches. We will try to improve the lower bound of the Skewes' number to the 27th power of 10 with a computational statistics approach, in which the number of prime number is statistically counted with the estimated errors by use of the sieve of Eratosthenes.

CP0572: **Interval and value at risk forecasting for realized volatility and implied volatility having asymmetry**
Presenter: **Ji Eun Choi**, Ewha Womens university, Korea, South
Co-authors: Dong Wan Shin

A new strategy is proposed for forecasting the confidence interval (CI) and value at risk (VaR) of Realized volatility (RV) and implied volatility (IV) which fully addresses asymmetry. For all RVs and IVs, significant asymmetries are identified in three parts; mean, volatility and error distribution. No method for asymmetries reflected multi step CI and VaR forecasts are available in the literature. The asymmetries are addressed by the LHAR (Leverage heteroscedastic autoregression) model for mean part, by the EGARCH model for volatility part, by the skew-t distribution for residual

part. Considerable out-of-sample forecast improvements of the CI and VaR is demonstrated for three financial assets: the US S&P 500 index, the US NASDAQ index, the Korea KOSPI index. For the RVs and IVs, we get CI forecasts with better coverage with smaller length and better VaR with better violation error if asymmetries are properly considered.

CP0424: **The limiting distribution of a rank test based on the multisample Cucconi test**

Presenter: **Takuya Nishino**, Tokyo University of Science, Japan

Co-authors: Hidetoshi Murakami

Various test statistics for the two-sample location-scale problem have been proposed. A nonparametric one-way layout analysis of variance plays an important role in biometry. We consider a multisample nonparametric statistic for the location-scale problem. The multisample Cucconi test is suitable for the shifted location-scale parameter. However, the limiting distribution of the multisample Cucconi test has not been derived. We derive the limiting distribution of the multisample Cucconi test. We investigate the convergence of the multisample Cucconi test to the limiting distribution for various cases by simulation studies. We estimate the exact critical value by a permutation method in which includes 1,000,000 replications.

CC0213: **Regression on a unit sphere**

Presenter: **Jayant Jha**, Indian Statistical Institute, India

Co-authors: Atanu Biswas

Spherical data is not very well-studied in statistics, although there are numerous works on linear random variables and linear regression. A spherical random variable is concerned with the orientation of points in k-dimensional space, and generally spherical data is presented through the points on the surface of a unit sphere. The work is an attempt to address the regression problem on a finite dimensional unit sphere when both the covariate and the response variables are spherical in nature. The model is explained first. Then, some distributional and geometric properties related to the model are studied. The algorithm is shown for generating data from such a setup when the error follows exit distribution. This algorithm is based on Brownian motion. The inferential issues are also addressed and comparison with existing models are discussed. A brief discussion on how to compute the MLE when the error follows von Mises Fisher distribution is also done. Some real datasets are used to illustrate the proposed method.

Wednesday 24.08.2016 14:30 - 16:00 Parallel Session H – COMPSTAT

COMPUTATIONAL CHALLENGES IN EXTREMES

CI085 **Room Sala Camara** **Chair: Marc Genton**

CI0210: **High-order composite likelihood inference for max-stable distributions and processes**
Presenter: **Stefano Castruccio**, Newcastle University, United Kingdom

In multivariate or spatial extremes, inference for max-stable processes observed at a large collection of points is a very challenging problem, and current approaches typically rely on less expensive composite likelihoods constructed from small subsets of data. We explore the limits of modern state-of-the-art computational facilities to perform full likelihood inference and to efficiently evaluate high-order composite likelihoods. With extensive simulations, we assess the loss of information of composite likelihood estimators with respect to a full likelihood approach for some widely used multivariate or spatial extreme models, we discuss how to choose composite likelihood truncation to improve the efficiency, and we also provide recommendations for practitioners.

CI0325: **Censored local likelihood inference for modeling non-stationarity in spatial extremes**
Presenter: **Daniela Castro**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Raphael Huser

In order to model complex dependence structures in spatial extremes, we propose an approach based on factor copula models. The latter, which can be seen as Gaussian location mixture processes, assume the presence of a common factor affecting the joint dependence of all measurements. When the common factor is exponentially distributed, the resulting copula is asymptotically equivalent to the Husler-Reiss copula; therefore, the so-called exponential factor model is suitable to capture tail dependence. Under the assumption of local stationarity, the exponential factor model is used to model non-stationary extreme measurements over high thresholds. Inference is performed using a censored local likelihood. Performance is assessed using simulation experiments, and illustrated using a daily rainfall dataset.

CI0331: **The use of Bernstein polynomials for modelling the extremal dependence**
Presenter: **Simone Padoan**, Bocconi University, Italy

A simple approach for modelling multivariate extremes is to consider the vector of component-wise maxima and their max-stable distributions. The extremal dependence can be inferred by estimating the angular measure or, alternatively, the Pickands dependence function. A flexible means for modelling such a dependence structure is the use of polynomials in the Bernstein form. For example, in the bivariate case, we describe a simple nonparametric Bayesian model that allows the estimation of both functional representations, satisfying the constraints required in order to provide a valid extremal dependence. This task is attained by placing a prior distribution on the Bernstein polynomials' coefficients, which gives probability one to the set of valid functions. The prior is extended to the polynomial degree, making our approach fully nonparametric. We show how to infer the extremal dependence, represented by Bernstein polynomials, using also the frequentist approach. With both approaches our proposal turns out to be a flexible framework for estimating component-wise maxima as well as the threshould exceedances. We show the utility of our proposed methods by a simulation study and a real data analysis.

COPULA REGRESSION

CO051 **Room Sala 1** **Chair: Thomas Kneib**

CO0158: **Some comments on copula-based regression**
Presenter: **Ria Van Hecke**, Ruhr Universitaet Bochum, Germany
Co-authors: Holger Dette, Stanislav Volgushev

A new semiparametric estimate of a regression function with a multivariate predictor was recently proposed, which is based on a specification of the dependence structure between the predictor and the response by means of a parametric copula. We investigate the effect which occurs under misspecification of the parametric model. We demonstrate by means of several examples that even for a one-, or two-, dimensional predictor, the error caused by a wrong specification of the parametric family is rather severe if the regression is not monotone in one of the components of the predictor. Moreover, we also show that these problems occur for all of the commonly used copula families, and we illustrate in several examples that the copula-based regression may lead to invalid results even when flexible copula models such as vine copulae (with the common parametric families) are used in the estimation procedure.

CO0159: **Simultaneous inference in structured additive conditional copula regression models: A unifying Bayesian approach**
Presenter: **Thomas Kneib**, University of Goettingen, Germany
Co-authors: Nadja Klein

While most regression models focus on explaining distributional aspects of one single response variable alone, interest in modern statistical applications has recently shifted towards simultaneously studying multiple response variables as well as their dependence structure. A particularly useful tool for pursuing such an analysis are copula-based regression models since they enable the separation of the marginal response distributions and the dependence structure summarised in a specific copula model. However, so far copula-based regression models have mostly been relying on two-step approaches where the marginal distributions are determined first whereas the copula structure is studied in a second step after plugging in the estimated marginal distributions. Moreover, the parameters of the copula are mostly treated as a constant not related to covariates and most regression specifications for the marginals are restricted to purely linear predictors. We therefore propose simultaneous Bayesian inference for both the marginal distributions and the copula using computationally efficient Markov chain Monte Carlo simulation techniques. In addition, we replace the commonly used linear predictor by a generic structured additive predictor comprising for example nonlinear effects of continuous covariates, spatial effects or random effects and furthermore allow to make the copula parameters covariate-dependent.

CO0273: **Estimating distribution functions using double sampling designs**
Presenter: **Ori Davidov**, University of Haifa, Israel

In some situations it is either infeasible or too expensive to collect data on the true outcome of interest from all sampling units. If an easy to collect proxy for the true outcome exists, then a double sampling design may be beneficial. In such situations there typically exists a large primary sample on which the proxy data is available for all sampling units. On a subsample of units from the primary sample the true outcome is ascertained. This subsample is commonly referred to as the validations sample. The full data is then used to estimate the parameter of interest. We propose two methods for empirically estimating a distribution function, and consequently its functionals, under double sampling. The first method is completely nonparametric and the second method is semiparametric copula based. Theoretical properties of the proposed estimators are investigated and simulation experiments presented. Extensions to multivariate distributions, conditional distributions and censored data and high dimensional data are discussed. The methodology is illustrated using a real data example.

CO0282: **Bivariate copula additive models for location, scale and shape**
Presenter: **Giampiero Marra**, University College London, United Kingdom
Co-authors: Rosalba Radice

A unified framework is discussed for fitting flexible bivariate copula-based regression models for continuous margins, binary margins and a mixture

of the two. The proposed approach allows for the simultaneous estimation of the copula coefficient and marginal distribution parameters (typically location, scale and shape), and for each parameter to be modelled in a regression setting using an additive predictor that comprises different types of covariate effects (e.g. non-linear, random and spatial effects). Parameter estimation is achieved within a penalized likelihood framework using a computationally stable and efficient trust region algorithm with integrated automatic multiple smoothing parameter selection. The proposed approach allows for straightforward inclusion of potentially any parametric marginal distribution and copula function. The models can be easily used via the R package SemiParBIVProbit. The usefulness of the proposal will be illustrated on several case studies drawn from the fields of epidemiology, biostatistics and econometrics.

ADVANCES IN COMPUTATIONAL STATISTICS AND STATISTICAL MODELLING II

CO108 **Room Sala 3** **Chair: Inmaculada Barranco-Chamorro**

CO0479: **Combination of forecasts in dynamic factor models: Application to the Italian power exchange**
Presenter: **Carolina Garcia-Martos**, Universidad Politecnica de Madrid, Spain
Co-authors: Andres M Alonso, Guadalupe Bastos

Nowadays, electricity markets are liberalized in the vast majority of countries of our socio-economic context. That is why forecasting electricity prices is a crucial task, both in the short (one day up to a week ahead) and long run (year-ahead). In the last decade, many methodological contributions have been developed in order to cope with empirical features of electricity price series. Particularly, Dynamic Factor Models (DFM) have emerged as a very interesting alternative for reducing forecasting errors both in the short and long run. However, when estimating a DFM, a model for the common factors must be selected. This task can be difficult, and it can occur that there is no a unique valid model for each factor. We propose model averaging as a possible alternative to deal with this issue, expecting that this would give smaller forecasting errors than modelling common factors just by a single model. Numerical results are provided for electricity spot prices in the Italian Power Exchange (IPEX) but the presented methodology can be applied to any other market. Results of combinations of forecasts are shown for different forecasting horizons.

CO0462: **The expectation maximization algorithm for the state space model with correlated errors**
Presenter: **Javier Cara**, Universidad Politecnica de Madrid, Spain

The state space model is a well known and used time series model, specially when dealing with multivariate time series. In practice, this model can be estimated using the Expectation Maximization algorithm assuming both the error in the state equation and the error in the observation equation are not mutually correlated. However, there are some situations where this hypothesis is not valid. The equations for estimating the state space model with correlated errors using the Expectation Maximization algorithm are presented. Finally, these equations are applied to time series of acceleration data measured in civil engineering structures, one example of multivariate state space model with correlated errors.

CO0463: **Important variable assessment in modeling transport accident patterns using random forest and bagging**
Presenter: **Camino Gonzalez**, Technical University of Madrid, Spain
Co-authors: Blanca Arenas, Belen Jimenez

Phenomenology related to traffic accident is really complex: many factors and variables involved, and usually, data recorded from police reports are heterogeneous and contain missing values. Besides relationships between variables may be strongly non linear, involve some high order interactions and quantitative variables are far for being normally distributed. Under these conditions prediction is very difficult, and the commonly used statistical modeling techniques sometimes are not enough efficient to display meaningfully the underlying accident pattern. Random Forest and Bagging address both exploring and modeling complex data base and play an important role to discover hidden relation between variables, also complemented with high prediction accuracy. The data base used includes transport accidents in the Spanish interurban roads from 2010 to 2012 (90000 records, 60 variables each). Predictor variables comprise functional road type, accident scenario characteristics, calendar variables and traffic information. The fitted models are used to define patterns associated to run-off-road and frontal collisions and also to create parsimonious models with high prediction accuracy. Additionally, several computational experiments have been conducted to perform sensitivity analysis on tuning parameters of algorithms (implemented in R and Guide) and to discuss on the appropriateness and effectiveness of the importance metrics.

CO0166: **Estimating extra zeros proportion in different variability conditions of a regression count model**
Presenter: **Antonio Jose Saez-Castillo**, Universidad de Jaen, Spain
Co-authors: Antonio Conde-Sanchez, Ana Maria Martinez-Rodriguez

Several identification problems in the fit of a count data model appear when the dataset includes a certain proportion of extra-zeros and is additionally affected by a lack of equi-dispersion: commonly, the zero inflation may be confused with over-dispersion, and under-dispersion may be also hidden if the model is not adequate. The zero inflated hyper-Poisson regression model permits to independently manage the existence of extra zeros and over- and/or under-dispersion in presence of covariates. A simulation study has shown its capability to adequately estimate the proportion of extra-zeros and to distinguish the existence of over- and under-dispersion. This model is tested in different real datasets which present different variability structure, and its fits are compared with those provided by other common models, such as the zero-inflated Poisson or the zero-inflated negative binomial.

ORDINAL AND CATEGORIAL DATA

CO010 **Room Sala 4** **Chair: Jae Chang Lee**

CC0522: **The analysis of asymmetry based on ordered scores for square contingency tables**
Presenter: **Hiroyuki Kurakami**, Tokyo University of Science, Japan
Co-authors: Shuji Ando

For the analysis of square contingency tables with same row and column ordinal classifications, the symmetry (S) model has been considered. The S model indicates that the cell probabilities of the table are symmetric with respect to the main diagonal line in a square contingency table. As an extension of the S model, some models having the structure of asymmetry based on the integer scores has been considered. We focus on the asymmetry models based on the ordered scores. We propose a new model based on the ordered scores. The proposed model indicates that the log odds of the symmetric cell probabilities decreases depending on the difference between the ordered scores. Also, we give the decomposition of the S model using the proposed model, and show that the test statistic for the S model is approximately equivalent to the sum of those for decomposed models. Additionally, the simulation study about the relationship between the new model and continuous distributions is given.

CC0455: **On separation of symmetry for ordinal categorical data**
Presenter: **Kouji Tahata**, Tokyo University of Science, Japan

Let us consider some trials which have more than two possible outcomes with ordinal categories. Then, the counts have the multinomial distribution. A problem of modeling of asymmetry for multinomial parameters with respect to the midpoint of categories is treated. Also, it is proved that the symmetry of multinomial parameters can be separated into the asymmetric structure of multinomial parameters and the structure of moment. This result can be applied for the multi-way contingency tables. For example, the decomposition of the point-symmetry mode was given time ago. Using these results, we can obtain a more detail decomposition of point-symmetry.

CC0528: **Entropy based tests of dependence for categorical data**
Presenter: **Simone Giannerini**, University of Bologna, Italy

Co-authors: Greta Goracci

The literature on measures of dependence and on tests for independence for categorical data is very wide. We focus on the power-divergence family of statistics, that includes as a special case both the measure S_ρ and Pearson's chi square. Nevertheless, the theoretical derivations concern the null hypothesis of stochastic independence against the alternative of some sort of local deviation from it. We derive an asymptotic approximation valid for the whole family and for every given level of dependence. This allows to build tests where $H_0 : S_\rho = S_0 \geq 0$ against $H_1 : S_\rho \neq S_0$. Moreover, we can compute analytically the power of the tests for independence that rely on the power-divergence family. We compare the performance of tests based on S_ρ with that of classical χ^2 both analytically and by means of Monte Carlo studies.

CC0363: **EMcorrProbit R package**

Presenter: **Denitsa Grigorova**, Sofia University, Bulgaria

Co-authors: Nina Daskalova

Correlated probit models (CPMs) are widely used for modeling of ordinal data or joint analyses of ordinal and continuous data which are common outcomes in medical studies. When we have clustered or longitudinal data CPMs with random effects are used to take into account the dependence between clustered measurements. When the dimension of the random effects is large, finding of the maximum likelihood estimates (MLEs) of the model parameters via standard numerical approximations is computationally cumbersome or in some cases impossible. EM algorithms for one ordinal longitudinal variable and for one ordinal and one continuous longitudinal variable are recently developed. The methods developed set the foundations of the EMcorrProbit R package (https://github.com/ninard/EMcorrProbit) which is going to offer also MLEs of CPM for two longitudinal ordinal variables via recently developed ECM algorithm. An application of the algorithm is presented to CPM for the longitudinal ordinal outcomes self-rated health and categorized body mass index from the Health and Retirement Study. We report results from fitting the model and also some simulation studies.

APPLIED ECONOMETRICS AND FINANCE

CG032 **Room Sala 7** **Chair: Stephen Pollock**

CC0201: **Globalization and economic growth: Tracking the impacts of changes in global interdependences**

Presenter: **Maria Jesus Delgado Rodriguez**, Universidad Rey Juan Carlos I de Madrid, Spain

Co-authors: Sonia de Lucas Santos

The research employs cyclical convergence analysis and simulation in panel data methodology to evaluate the potential effects of globalization in economic growth. On the basis of this approach, we derive alternative scenarios depending on different assumptions between global interdependences in the regions of the world. Results show that the dynamics of globalization in the last decades have been more driven by convergence in regional growth patterns than by the synchronization of the world economy as a whole.

CC0564: **Functional impact into Google AdWords**

Presenter: **Christoph Rust**, University of Regensburg, Germany

One of the core challenges of the online advertising industry is to explain sales conditional on various attributes of sponsored search advertisements (such as clicks, impressions, ranking, length of ad words). Functional regression with its typically involved dimension reduction techniques seems to provide a very suitable approach. Though, by contrast to classical functional regression problems, we need to focus not only on the functional explanatory variables, but also on specific point-wise information. For instance, time-point specific market events can have important effects that are contrary to the general temporal market evolution. To be able to capture both effects, we adopt a recent method for functional linear regression. Specifically, we analyze a big data sample of an AdWords retailer and propose a functional model for the whole dependence structure, from auction positioning to expected clicks and sales. Furthermore, we interpret findings resulting from the above mentioned method and compare the performance with common models.

CC0503: **yuimaGUI: A graphical user interface for computational finance based on the yuima R package**

Presenter: **Emanuele Guidotti**, University of Milan, Italy

Co-authors: Stefano Iacus, Lorenzo Mercuri

The aim of Yuima project is to develop a complete environment for simulation and inference of Stochastic Differential Equations (SDE) via an R package called yuima. The package is developed using the object oriented programming language S4 and allows the user to manage a stochastic process characterized by a SDE with the following general form: $dXt = b(t,Xt,\theta)dt + a(t,Xt,\theta)dWHt + c(t,Xt,\theta)dZt$, where $b(t,Xt,\theta)$, $a(t,Xt,\theta)$ and $c(t,Xt,\theta)$ are functions defined by the user. WH is the fractional Brownian motion where the Hurst coefficient H is fixed by default to 1/2 that corresponds to the standard Brownian motion.The yuima package has also estimation and simulation routines for two models widely used in modern computational finance: the Continuous ARMA(p,q) driven by a general Levy process and the COGARCH(p,q) model for high frequency data. Moreover, yuimaGUI allows to simulate a portfolio of assets and derivatives under different scenarios and further evaluate thee overall returns along with their distribution. After a brief presentation of the yuima framework, we will focus on aspects related to computational finance and show how they can be easily approached through the yuimaGUI package. We will go through data mining and clustering of financial time series, model selection, change point estimation, scenario simulation and the analysis of the distribution of the expected returns for composite portfolios.

CC0517: **Automatic ARIMA modeling using RcmdrPlugin.SPSS**

Presenter: **Dedi Rosadi**, Universitas Gadjah Mada, Indonesia

In some application of time series modeling, it is necessary to obtain forecast of various types of data automatically and possibly, in real-time way, for instance, to do a real-time processing of the satellite data. Various automatic algorithms for modeling ARIMA models are available in the literature, where we will discuss two methods in particular. One of the method is based on a combination between the best exponential smoothing model to obtain the forecast, together with state-space approach of the underlying model to obtain the prediction interval. Other method, which is more advanced method, is based on X-13-ARIMA-SEATS, the seasonal adjustment software by the US Census Bureau. These approaches are implemented in our R-GUI package RcmdrPlugin.Econometrics which now already integrated in our new and more comprehensive R-GUI package RcmdrPlugin.SPSS. We provide application of the methods and the tool using real data.

CLUSTERING

CG005 **Room Sala 5** **Chair: Gunter Ritter**

CC0410: **Variable importance in clustering using binary decision trees**

Presenter: **Pierre Michel**, Aix Marseille University, France

Co-authors: Badih Ghattas

Different approaches are considered for assessing variable importance in clustering. We focus on clustering using binary decision trees, which is a non-parametric top-down hierarchical clustering method designed for both continuous and nominal data. We suggest a measure of variable importance for this method similar to the one used in Breiman's classiffcation and regression trees. We analyze the effciency of this score on different data simulation models in presence of noise, and compare it to other classical variable importance measures.

CC0526: **Automatic module selection from several microarray gene expression studies**

Presenter: **Alix Zollinger**, SIB, Switzerland

Co-authors: Anthony Davison, Darlene Goldstein

Independence of genes is commonly but incorrectly assumed in microarray data analysis; rather, genes are activated in co-regulated sets referred to as modules. We develop an automatic method to define modules common to multiple, independent studies. We use an empirical Bayes procedure to estimate a sparse correlation matrix for all studies, identify modules by clustering, and develop an extreme-value-based method to detect so-called scattered genes, which do not belong to any module. The resulting algorithm is very fast and produces accurate modules in simulation studies. Application to real data identifies modules with significant enrichment and results in a huge dimension reduction, which can alleviate the computational burden of further analyses.

CC0252: **Comparison of two bootstrap procedures in the case of hidden Markovian model clustering**

Presenter: **Zhivko Taushanov**, University of Lausanne, Switzerland

Co-authors: Andre Berchtold

The clustering of longitudinal data sequences is considered using a latent Markovian model (HMTD) combining Gaussian distributions and covariates. The main objective is to evaluate the significance of the estimated parameters. At first, different model specifications are optimized and the one providing the best clustering in terms of BIC is selected. Two different bootstrap procedures are then applied and compared in order to investigate the significance of the parameters of this optimal solution. First, a standard bootstrap procedure is applied using the full original sample and the optimal model with multiple components (clusters) is computed at each iteration. That leads to solutions with different degrees of similarity with the optimal solution and the well-known label-switching problem may occur. An alternative procedure is proposed that consists in applying separate bootstrap procedures on each subsample defined by the optimal clustering. In this case, a single component model is estimated from each bootstrap iteration and for each cluster separately. This method also provides a confidence interval for each parameter and avoids the label-switching problem. The pros and cons of each approach are described and examples based on real data are provided.

CC0340: **Determining the number of overlapping clusters: Simulation results for the additive profile clustering model**

Presenter: **Tom Frans Wilderjans**, Leiden University, Netherlands

Co-authors: Julian Rossbroich

In various fields of science, researchers are interested in revealing the underlying structural mechanisms that generated object by variable data (e.g., patient by symptom or consumer by brand data). Based on theoretical or empirical arguments, it may be hypothesized that these underlying mechanisms are captured by a clustering of the objects. To this end, researchers often adopt a partitioning method (e.g. *k*-means), which yields a set of non-overlapping clusters. However, in some cases it may be expected that objects are grouped into clusters that are allowed to overlap (i.e. an object belonging to multiple clusters). For the patient by symptom data it may, for example, be that the clusters correspond with syndromes and that patients may suffer from multiple syndromes at the same time (i.e. co-morbidity). To identify the overlapping object clusters, Mirkins additive profile (overlapping) clustering model may be used. A non-trivial task consists of determining the optimal number of overlapping clusters underlying a data set at hand. Up to now, however, this issue of model selection has not been studied in a systematic way. Therefore, we compare in an extensive simulation study various existing (e.g. CHull, cross-validation) and new model selection methods for the additive profile clustering model, with some of the new methods being methods for the partitioning case tailored to the context of overlapping clustering (e.g. AIC, CH-index).

ROBUST STATISTICS II

CC071 **Room Sala 2** **Chair: Peter Rousseeuw**

CC0532: **On the computation of symmetrized M-estimators of scatter**

Presenter: **Jari Miettinen**, University of Jyvaskyla, Finland

Co-authors: Klaus Nordhausen, Sara Taskinen, David Tyler

The symmetrized version of a multivariate scatter functional is obtained when the functional is applied to the pairwise differences of the observations. Sometimes the symmetrization increases the efficiency, but perhaps the most important benefit is that the symmetrized scatter matrices always have the independence property, that is, they are diagonal when the components are mutually independent. A scatter matrix needs the independence property when it is used as a robust substitute of the sample covariance matrix in certain multivariate methods, as for example in independent component analysis and graphical modeling. The obvious disadvantage of symmetrization is the number of pairwise differences which grows very fast with the number of observations. The computational aspects of symmetrized M-estimates are considered. The computation time of the standard fixed point algorithm is compared to those of incomplete, parallel, and partial Newton algorithms. Also the efficiency loss of the incomplete estimator, which uses only a subset of the pairwise differences, is studied.

CC0543: **Two-step robust estimation of copulae**

Presenter: **Samuel Orso**, University of Geneva, Switzerland

Co-authors: Stephane Guerrier, Maria-Pia Victoria-Feser

Copula is a flexible tool for modeling multivariate random variables. Inference is generally based on multi-step estimators to preserve this flexibility. We address the problem of robustness in this context. It is challenging for many reasons: (a) How to built a gross error model that encompasses issues arising in multiple dimensions? (b) How to built multi-step robust estimators with good asymptotic properties? (c) How to obtain computationally feasible estimators and their inference? We concentrate our efforts in the two-dimensional case. First, we propose a new gross error model from which the influence function is derived for two-step M-estimators. Second, we prove the strong consistency and asymptotic normality under weak conditions. Third, we propose a fast bootstrap procedure to obtain the covariance matrix of the two-step estimators. We illustrate the estimating procedure with a new R package under development.

CC0474: **On a modification of Efron bootstrap method for heavy-tailed distributions**

Presenter: **Hannah Opayinka**, University of Ibadan, Nigeria

The nature of the upper tail of a heavy-tailed distribution is the major reason for the poor performance of classical bootstrap methods, which results in large standard error. Efron's bootstrap method is modified to address the challenges faced when dealing with heavy-tailed distributions. The methodology involves stratifying the observations into homogenous subgroups and using proportional allocation method in selecting bootstrap samples. Real and simulated data sets drawn from Lognormal and Singh-Maddala distributions respectively were used. The two distributions have finite variance. The performance of the Modified Efron Bootstrap (MEB) is compared to that of Efron's bootstrap using SE (standard error) and RMSE (root mean squared error). The findings show that SEs and RMSEs for all sample sizes considered were consistently smaller in MEB than Efron bootstrap. Hence, MEB outperformed Efron's bootstrap when applied to heavy-tailed distributions for the two cases considered.

CC0391: **Evaluation of robust PCA for supervised audio outlier detection**

Presenter: **Sarka Brodinova**, Vienna University of Technology, Austria

Co-authors: Thomas Ortner, Peter Filzmoser, Maia Zaharieva, Christian Breiteneder

Outliers often reveal crucial information about the underlying data such as the presence of unusual observations that require for in-depth analysis. The detection of outliers is especially challenging in real-world application scenarios dealing with high-dimensional and flat data bearing different subpopulations of potentially varying data distributions. In the context of high-dimensional data, PCA-based methods are commonly applied to reduce dimensionality and to reveal outliers. Thus, a thorough empirical evaluation of various PCA-based methods for the detection of outliers in

a challenging audio data set is provided. The various experimental data settings are motivated by the requirements of real-world scenarios, such as varying number of outliers, available training data, and data characteristics in terms of potential subpopulations.

Thursday 25.08.2016 09:00 - 10:30 Parallel Session I – COMPSTAT

APPLIED FUNCTIONAL DATA ANALYSIS

CI081 **Room Sala Camara** **Chair: Juan Romo**

CC0547: **Directional depth and outlyingness for multivariate functional data**
Presenter: **Marc Genton**, KAUST, Saudi Arabia
Co-authors: Wenlin Dai

The direction of outlyingness is crucial to describing the centrality of multivariate functional data. Motivated by this idea, we propose a new framework that combines classical depth with the direction of outlyingness. We generalize classical depth/outlyingness to directional depth/outlyingness in both point-wise and functional data. We investigate the affine invariance of directional functional depth and find that it naturally decomposes functional depth into two parts: a scale depth and a shape depth, which represent the centrality of a curve for magnitude and shape, respectively. Using this decomposition, we provide a visualization tool for the centrality of curves. Furthermore, we design an outlier detection procedure based on directional functional outlyingness. This criterion applies to both univariate and multivariate curves and simulation studies show that it outperforms existing methods. Weather and electrocardiogram data demonstrate the practical application of our proposed framework.

CI0347: **Analyzing high-dimensional functional data**
Presenter: **Juan Romo**, Universidad Carlos III de Madrid, Spain
Co-authors: Ana Arribas-Gil

Functional data not only arrive as one-dimensional collections, but also as multivariate samples of curves or even as high- dimensional functional data sets. The concept of depth provides convenient tools for the robust analysis of curves. It allows to establish the notion of centrality and extremality and provides fundamentals for testing or classification. We analyze depth for high-dimensional functional data and apply these ideas to simulated high-dimensional samples of curves and to real high-dimensional functional observations.

CI0576: **Bayesian multivariate spatial temporal functional data modelling**
Presenter: **Montserrat Fuentes**, North Carolina State University, United States

It is well known that the volume and complexity of scientific data is increasing. This increase necessitates the development of flexible, and efficient, statistical methods which are capable of accurately capturing this complexity. Motivated by the analysis of hurricane trajectories and intensities we develop a Bayesian, multivariate, functional linear model with spatially varying coefficients. The model utilizes a hierarchical structure in order accommodate noisy functional covariates and allow for the inclusion of derivatives of functional covariates. In addition, tensor product basis expansions paired with appropriately structured prior distributions are used to allow for spatially adaptive coefficients. Temporal correlation within storms is modeled using an autoregressive term. Appropriate specification of the prior distributions allows Gibbs sampling to be used for the entire model. Posterior inferences are then constructed using MCMC output.

ADVANCES IN THE PATTERN RECOGNITION OF TIME SERIES

CO027 **Room Sala 1** **Chair: Elizabeth Ann Maharaj**

CO0153: **Pattern recognition techniques for interval time series**
Presenter: **Elizabeth Ann Maharaj**, Monash University, Australia
Co-authors: Paula Brito, Paulo Teles

An interval time series (ITS) is a sequence of intervals observed in successive instants in time. We focus on the clustering of a set of ITS where we examine existing techniques and propose some new techniques. One new technique involves using a measure that determines the degree of overlap of every pair of ITS under consideration. The measure lies between zero and one and the closer it is to one, the greater the degree of overlap of the two ITS. A distance-type matrix consisting of these measures is used as the input into hierarchical clustering methods. Another new technique involves fitting space-time models to each of the ITS under consideration, and using the parameter estimates of the fitted models as inputs into newly developed and existing clustering methods. Where these techniques differ from existing ones which take into account the upper and lower bounds individually, is that the link between the upper and lower bounds is taken into account.

CO0255: **Adaptive spectral analysis of replicated nonstationary time series**
Presenter: **Robert Krafty**, University of Pittsburgh, United States
Co-authors: Scott Bruce, Martica Hall

A new method is discussed for analyzing associations between nonstationary time series and cross-sectional variables when data are observed from replicated independent units or subjects. The approach adaptively divides time series and values of the cross-sectional variable into approximately stationary blocks, then estimates conditional local power spectra nonparametrically through Whittle likelihood based smoothing splines. The model is formulated in a Bayesian framework and fit via reversible jump MCMC methods, which allow for the modeling of both abrupt and smoothly varying effects. The method is used to analyze data from a study of caregivers of spouses with dementia and uncovers connections between heart rate variability during sleep and quality of life.

CO0292: **Quantile autocovariances: A powerful tool for hard and soft partitional clustering of time series**
Presenter: **Jose Vilar**, Universidade da Coruna, Spain
Co-authors: Borja Lafuente-Rego

A distance based on estimated quantile autocovariances (QAD) is proposed to perform time series clustering when the target is to group series according to the underlying dependence structures. Unlike other extracted features, quantile autocovariances account for sophisticated dynamic features and are well-defined for a broad class of processes. Hence, a cluster procedure based on comparing quantile autocovariances should report satisfactory results, particularly in complex scenarios involving non-linear or heteroskedastic processes. The behavior of QAD in partitioning-based clustering is examined considering both crisp and fuzzy procedures. Contribution consists of three points. First, a broad simulation study shows the good behavior of the QAD-based clustering compared to other commonly used dissimilarities. Excellent scores are attained by classifying heteroskedastic processes and also when non-normal innovations are considered. Second contribution concerns the optimal selection of input parameters, i.e. the problem of determining the proper combination of lags and quantile levels is addressed. Third contribution consists in introducing a novel fuzzy procedure based on QAD, which presents high capability to clustering GARCH models, outperforming fuzzy clustering algorithms specifically designed to work in this framework. The usefulness of the proposed procedure is illustrated by its application to a case-study.

OPTIMAL DESIGNS FOR COMPLEX MODELS VIA SIMULATION

CO039 **Room Sala 2** **Chair: Jesus Lopez-Fidalgo**

CO0228: **Optimal designs for fractional polynomial models**
Presenter: **Victor Casero-Alonso**, University of Castilla-La Mancha, Spain
Co-authors: Jesus Lopez-Fidalgo, Weng Kee Wong

Fractional polynomials have been shown to be much more flexible than polynomials for fitting continuous outcomes in the biological and health sciences. Despite their increasing popularity, design issues for fractional polynomials models have never been addressed. D- and I-optimal experimental designs for prediction using fractional polynomial models are provided, their properties are evaluated and a catalogue of design points useful for fractional polynomial models is provided. As applications, we re-design two studies using optimal designs and show they can produce substantial gains in terms of cost and quality of the statistical inference. We also provide a user friendly applet for generating optimal designs for fractional polynomials up to degree 3.

CO0307: **Bayesian optimal designs via MCMC simulations: A case study in the technological field**
Presenter: **Rossella Berni**, University of Florence, Italy

Optimal design criteria have recently received growing attention, both theoretically and computationally, also due to the increase of computational power. Since the 70s, there has been a long list of seminal papers about D and T optimality, both to estimate model parameters and to discriminate among models. Furthermore, optimal designs have been improved in a Bayesian framework, by introducing prior distributions on models and parameters and by selecting the optimal design according to the definition of an utility function and its maximization, in a decision analysis framework. Despite the generality achieved, in actual applications further flexibility is often needed: for example, when defining a utility function in which the cost of each observation depends on the value of the independent variables; also, the relevance for costs may be also evaluated by specific weights, which take environmental conditions and technological information into account. We improve optimal designs in the technological field by applying Markov Chain Monte Carlo simulations, and by evaluating: i) a hierarchical structure of the observed data; ii) an utility function including costs and weights; iii) model discrimination.

CO0421: **Optimal-design search under the IMSPE objective**
Presenter: **Selden Crary**, Unaffiliated, United States

Searches for optimal statistical designs of computer experiments, under the integrated mean-squared prediction error (IMSPE) objective, are often thought to encounter insurmountable problems because of ill-conditioning of the covariance matrix, whenever two or more trial design points are proximal in the design domain. The customary resolution is to disallow proximal design points, but doing so can disallow optimal designs that were the goal of the search. An alternative approach is to recognize the IMSPE is a member of a special class of pole-free, low-degree-truncated rational functions with essential discontinuities. Examples are provided of how optimal-design searches can be completed without excluding proximal points and how optimal designs with proximal points can be successfully used for metamodel generation.

CO0422: **Using sparse kernels to design computer experiments with tunable precision**
Presenter: **Guillaume Sagnol**, ZIB Berlin, Germany
Co-authors: Hans-Christian Hege, Martin Weiser

Statistical methods to design computer experiments usually rely on a Gaussian process (GP) surrogate model, and typically aim at selecting design points (combinations of algorithmic and model parameters) that minimize the average prediction variance, or maximize the prediction accuracy for the hyperparameters of the GP surrogate. In many applications, experiments have a tunable precision, in the sense that one software parameter controls the tradeoff between accuracy and computing time (e.g., mesh size in FEM simulations or number of Monte-Carlo samples). We formulate the problem of allocating a budget of computing time over a finite set of candidate points for the goals mentioned above. This is a continuous optimization problem, which is moreover convex whenever the tradeoff function accuracy vs. computing time is concave. On the other hand, using non-concave weight functions can help to identify sparse designs. In addition, using sparse kernel approximations drastically reduce the cost per iteration of the multiplicative weights updates that can be used to solve this problem.

FACTOR ANALYSIS-BASED METHODS

CG086 **Room Sala 3** **Chair: Nathalie Villa-Vialaneix**

CC0200: **Multiway-SIR for longitudinal multi-table data integration**
Presenter: **Nathalie Villa-Vialaneix**, INRA, France
Co-authors: Valerie Sautron, Marie Chavent, Nathalie Viguerie

An extension of DUAL-STATIS to the sliced-inverse regression (SIR) framework is proposed to analyze multi-table datasets with respect to a numeric variable of interest. The method is designed to analyze the case where a data set $\mathbf{X}$, which corresponds to a set of p variables measured T times on the same n subjects is related to a real target variable $\mathbf{y}$, measured on the same n subjects. The approach is an exploratory method which aims at understanding the evolution of the relation between $\mathbf{X}$ and $\mathbf{y}$ through time. The method proceeds in two steps: 1) an inter-structure analysis studies the resemblance between the different time steps by computing similarities between estimates of the covariance of the mean of $\mathbf{X}_{..t}$ conditional to $\mathbf{y}$. Similarly to SIR, the conditional expectation is estimated by slicing the range of $\mathbf{y}$. The result of this analysis is a compromise covariance matrix Γ^c, which captures a compromise correlation structure of $\mathbb{E}(\mathbf{X}_{..t}|y)$ over t; 2) an intra-structure analysis which is a generalized PCA of the compromise. This second step results in graphical outputs which can be used to explore the covariance structure between variables and time steps conditional to $\mathbf{y}$. The method is illustrated on a real problem related to the consequences of a low calorie diet on obese persons in which the target variable of interest is the weight gain/loss.

CC0480: **Equating multidimensional IRT parameters when both common items and common persons are available**
Presenter: **Yoshinori Oki**, Tokyo Institute of Technology, Japan
Co-authors: Shin-ichi Mayekawa

Item Response Theory (IRT) is a set of stochastic models for psychological and educational tests. IRT is based on the idea that the probability of a correct response to an item is represented by a mathematical function of item and person parameters. In the application of IRT, calibrating the parameters of two or more tests is a critical issue, because it allows for comparisons between test scores, and common item design or common person design is often used for the equating. Multidimensional IRT (MIRT) is a sub-model of IRT; in MIRT it is assumed that more than two abilities have effects on the probability of a correct answer. MIRT has indeterminacy between item and person parameters same as factor analysis, and one can use equating methods for rotation of parameter matrixes. There is a case that both common item and common person design are available in MIRT. However, few studies have been conducted corresponding to the case. We propose the integration of common item and common person criteria, and utilize a rotation method in factor analysis focusing on both factor scores and factor loadings as a method for equating in MIRT.

CC0395: **EM estimation of a structural equation model**
Presenter: **Myriam Tami**, University of Montpellier, France

A new estimation method of a Structural Equation Model (SEM) is proposed. Contrasting with the classical SEM approach, our method is not based on the constrained estimation of the covariance structure of the data. It consists in viewing the Latent Variables (LV's) as missing data and using the EM algorithm to maximize the whole model's likelihood, which simultaneously provides estimators not only of the model's coefficients, but also of the values of LV's. Through a simulation study, we investigate how fast and accurate the method is, and eventually apply it to real environmental data.

CC0293: **Constructing a composite indicator for education monitoring**
Presenter: **Dovile Stumbriene**, Vilnius university, Lithuania
Co-authors: Rimantas Zelvys, Audrone Jakaitiene

The focus is on the construction of a composite indicator for the education monitoring. It is common awareness that socio-economic phenomena are complex and cannot be measured by a single descriptive indicator it should be represented with multiple dimensions. Phenomena such as education can be measured and evaluated by applying methodologies known as composite indicators. Methods are reviewed to create the education monitoring index that apply data treatment and normalization procedures, weighting and aggregation strategy, which assigns weights to the components when combining them and chooses a synthetic function. At the data weighting stage we used factor analysis and data envelopment analysis in order to discuss how the different approaches affect the results. In order to compare the different methodologies, the education monitoring index was calculated for Lithuania, Latvia, Estonia, United Kingdom, Finland and Germany over time using EUROSTAT data. The index was constructed following structural CIPO framework, which describes relationships between Input, Process and Output in education within a certain Context.

BAYESIAN METHODS II

CG042 **Room Sala 5** **Chair: Alastair Young**

CC0452: **On the first-order integer-valued bilinear model**
Presenter: **Isabel Pereira**, University of Aveiro, Portugal
Co-authors: Nelia Silva

The integer-valued bilinear INBL $(1,0,1,1)$ model is discussed. This class of models is particularly suitable for modeling processes which assume low values with high probability, but exhibiting, at the same time, sudden bursts of large values. However, although the likelihood function is based on convolutions of commonly used distributions, it has a very complex form. In order to overcome this difficulty it is proposed an approach based on saddlepoint techniques to estimate model parameters. It is carried out a simulation study to compare these results with the ones obtained through maximum likelihood method and Bayesian approach. Furthermore, the problem of predicting future observations from the classical and Bayesian approaches is analyzed. Since the evaluation of predictive performance and the suitability of the model are important issues, the probabilistic forecast is compared with the true data-generating distribution. This comparison is made using the Probability Integral Transform (PIT) and applied to real data sets of E.coli infections and meningitis cases.

CC0511: **Distance dependent Chinese restaurant process for spatio-temporal clustering of urban traffic networks**
Presenter: **Ashwini Venkatasubramaniam**, University of Glasgow, United Kingdom
Co-authors: Konstantinos Ampountolas, Ludger Evers

A novel Bayesian clustering method is presented for spatio-temporal data observed on a network. This method employs a Distance Dependent Chinese Restaurant Process (DDCRP) to incorporate the geographic constraints of the network. DDCRPs typically accommodate non-exchangeable distributions as a prior over partitions unlike the traditional Chinese restaurant Process. In addition, it does not exhibit the marginal invariance property and so one can capture the extent of the influence transferred from one node in the network to the next. We do not expect the DDCRP to fully capture the dependency structure of the data and thus a conditional auto-regressive model (CAR) is used to model the spatial dependency within a cluster. We will discuss different strategies for incorporating temporal dependency into a CAR-type model. Inference is carried out using a Metropolis within Gibbs sampler and we apply the model to cluster an urban traffic network, using occupancy data recorded at the junction level.

CC0382: **A Bayesian approach for the transformed gamma degradation process**
Presenter: **Fabio Postiglione**, University of Salerno, Italy
Co-authors: Massimiliano Giorgio, Maurizio Guida, Gianpaolo Pulcini

Very recently, a new degradation process, namely the transformed gamma (TG) process, has been proposed in the literature to describe Markovian degradation processes whose increments over disjoint intervals are not independent, so that the degradation growth over a future time interval can depend both on the current age and the current state (degradation level) of the unit. We propose a Bayesian estimation approach for such a process, that is based on prior information relative to the sign (positive or negative) of the correlation between the degradation increment and the current state or age of the unit. Several different prior distributions are then proposed, reflecting the knowledge of the analyst. A Markov Chain Monte Carlo technique, based on the adaptive Metropolis algorithm, is used for estimating the TG parameters and some functions thereof, such as the residual reliability of a unit, as well as for predicting future degradation growth. Finally, the proposed approach is applied to a real dataset consisting of wear measures of the liners of the 8-cylinder engine which equips a cargo ship.

CC0371: **Bayesian prediction of unobserved values for Type-II censored data for an inverse Weibull distribution**
Presenter: **Takeshi Kurosawa**, Tokyo University of Science, Japan
Co-authors: Tatsuya Kubota

Censoring problems often arise in survival time analysis. In survival time analysis, a Weibull distribution and an inverse Weibull distribution are often used. Hence, we focus on Type-II censored data for the inverse Weibull distribution. The aim is to derive posterior predictive distributions of unobserved values for Type-II censored data. These functions are given by integrals of conditional density functions and conditional survival functions over two hyper parameters. The conditional survival functions were also expressed by integral forms in previous studies. The integrals of the conditional survival functions are calculated by Monte Carlo integrations and it takes too much time to compute. Therefore, we derive the conditional survival functions in closed forms using a theory of order statistics and thereby reduce the computation cost. In addition, we calculate the predictive confidence intervals of unobserved values, coverage probabilities of unobserved values by using the derived posterior predictive survival functions and confirm the correctness of our derived functions.

QUALITY CONTROL

CG036 **Room Sala 7** **Chair: Benjamin Reiser**

CC0482: **Variable selection with pre-assigned roles and cost utility analysis for process monitoring**
Presenter: **Luan Jaupi**, CNAM, France

Methods for selecting a reduced number of relevant variables for process monitoring are proposed. We assume that a two-class system is used to classify the variables as primary and secondary based on different criteria. Then a double reduction of dimensionality is applied to select relevant primary variables that represent well the whole set of variables. The selection methodology uses external information and a cost-utility analysis to influence the selection process and compare one use of resources with other possible uses. The subset of relevant variables is selected in a manner that retains, to some extent, the structure and information carried by the full set of original variables. The advantages of the methods are illustrated with simulation results and a real application from automotive industry.

CC0426: **A new EWMA chart for monitoring the covariance matrix of a multivariate process based on dissimilarity index**
Presenter: **Longcheen Huwang**, National Tsing Hua University, Taiwan

An EWMA chart is proposed for monitoring the covariance matrix of a multivariate process based on the dissimilarity index of two matrices. The conventional charts for monitoring the covariance matrix of a multivariate process are either based on comparing the sum or the product or both of the eigenvalues of the estimated covariance matrix with those of the in-control covariance matrix. In contrast, the proposed new chart essentially monitor the covariance matrix by comparing the individual eigenvalues of the estimated covariance matrix with those of the in-control counterpart. We compare the performance of the proposed chart with that of the best existing chart in the multivariate normal process. Simulation results show that the proposed EMMA chart outperforms the best existing multivariate EWMA chart for monitoring the covariance matrix. Further, to guarantee

that the actual in-control average run length of the proposed chart is not less than the nominal one with a certain probability, we use a bootstrap re-sampling method to adjust the control limit of the proposed chart. Finally, we use an example to demonstrate the applicability of the proposed chart.

CC0584: **Web computing of robust methods in Acceptance Sampling for Weibull variables**
Presenter: **Miguel Casquilho**, University of Lisbon, Portugal
Co-authors: Elisabete Carolino, Rosario Ramos, Isabel Barao

Acceptance sampling (AS) is used to inspect the input or output of a process, mainly in manufacturing. A sampling plan is designed to determine a procedure that, if applied to a series of lots of a given quality and based on sampling information, leads to a specified risk of accepting or rejecting the lots. Classic AS by variables assumes Gaussian distribution, as treated in industry standards, which is sometimes an abusive assumption leading to wrong decisions. AS for variables with asymmetric and/or heavy tailed distributions is then a relevant topic. Specific AS plans are derived for the case of the Weibull distribution. As an alternative to these AS plans for the Weibull, traditional plans are used with robust estimators. The estimators are the sample median for location and a modified version of the sample standard deviation and the Total Range for scale estimates. The problem of determining AS plans by variables is addressed for the Weibull distribution with unknown parameters. The aim is to apply computing over the Web, through an application made available to any user, needing no software installation, in order to conduct a simulation study on these methods, about classical plans, specific plans and plans using the robust estimates for location and scale

CC0386: **A modified control chart for monitoring the multihead weighing process**
Presenter: **Alexander Pulido-Rojano**, Universidad Simon Bolivar, Colombia
Co-authors: Juan Carlos Garcia-Diaz

The modified control charts are used for monitoring and control of the manufacturing processes which are considered as six-sigma process, ensuring a probability of out-of-specification product acceptably small. The use of these charts is based on the idea that the cost of identifying and correcting special causes is much higher than the cost of off-target products. Therefore, the process mean is essentially acceptable as long as it is anywhere within the specification limits. These concepts have been applied to the packaging process in multihead weighers. The weight of the packed product, seen as the quality characteristic to be monitored, must be as close to a specified target weight and comply with applicable regulations. In order to design the modified control chart and comply with requirements for its implementation, the packaging process has been previously optimised and improved through a packaging strategy. The strategy seeks to reduce the variability in the selection of the total weight of the package and it is evaluated through a proposed packing algorithm. In this way, a set of numerical experiments were conducted to examine the solutions generated and which are subsequently monitored.

REGRESSION MODELS

CG030 **Room Sala 4** **Chair: Thomas Yee**

CC0537: **Bootstrap-based inference in generalized linear mixed models**
Presenter: **Daniel Antonio Flores Agreda**, Universite de Geneve, Switzerland
Co-authors: Eva Cantoni

The focus is on two topics related to the inference of random effects using bootstrap methods. On a first stage we address the problem of uncertainty estimation in prediction for random effects in mixed models as measured by the Mean Squared Error of the Empirical Predictor (MSEP). We propose a non-parametric algorithm for estimation of the MSEP based on the Generalized Bootstrap for Estimating Equations adapted for the Linear Mixed Models setting. We apply this procedure in the framework of Generalized Linear Mixed Models and the EBP as we illustrate the properties of our proposal with simulation studies. On a second stage, we discuss extensions addressing the problem of random effect selection. We discuss the implementation of a penalized version of the Generalized Bootstrap for Estimating Equations with a re-parametrization that allows the construction of bootstrap confidence intervals for fixed and random effect parameters.

CC0169: **On negative binomial regression and its variants**
Presenter: **Thomas Yee**, University of Auckland, New Zealand

The negative binomial distribution (NBD) is a very popular alternative to the Poisson for handling over-dispersion. We look at some computational issues relating to estimating the parameters of the NBD and some of its variants. These variants include the NB-1 (whose variance function is proportional to the mean), the NB-H (whose ancillary parameter is modelled using the same explanatory variables as the mean), and the NB-P (a more flexible data-driven NBD variant). It is shown that the statistical framework described recently is flexible enough to include all these models with surprising ease. Some results from the author's VGAM R package will be included.

CC0512: **Impact of correlation between predictors on variance decomposition and variable selection using CAR scores**
Presenter: **Henri Wallard**, Ipsos, France

Regression is widely used to identify the most important regressors or to quantify their relative importance. If predictors are strongly correlated OLS can be difficult to use and regularisation methods or variance decomposition have been considered. Whether the measures of importance computed tend to equal or not when a group of regressors are highly correlated has been investigated for regularisation methods like ridge regression or elastic net, but also for variance decomposition through CAR scores. CAR scores have been presented as providing a canonical ordering of importance and been credited with grouping property in the sense that CAR scores would become identical with growing value of the correlation between variables. It is demonstrated that CAR scores do not benefit from this grouping property. Using geometric interpretation and theoretical examples in the case of two and three predictors we will show in contrary that CAR scores remain stable or can even grow when the correlation between predictors increases, and confirm on real dataset that grouping property is not achieved with CAR scores leading to risk of inconsistency for variable selection. As such CAR scores for variable importance or selection are not recommended and other methods will be proposed.

CC0428: **On constrained estimation of graphical time series models**
Presenter: **Heung Wong**, The Hong Kong Polytechnic University, China

Graphical models represent the conditional independence relations between random variables in multivariate data. These independence relationships can be visualized by an undirected graph where vertices represent the variables and edges between vertices illustrate that the corresponding variables of the connected vertices are conditionally dependent. The increasing interest in data science has heightened the need for the development of Gaussian graphical models with sparse coefficients on high dimensional data. The use of graphical models has been extended to multivariate time series to explore the interrelationship between components of a multivariate time series process. We propose a method to estimate a graphical time series model, which is based on a sparse vector autoregressive process. Both the autoregressive coefficients and the entries of the inverse of the noise covariance matrix will be estimated. To impose sparsity on both the coefficients and the inverse covariance matrix, we propose an iterative algorithm to estimate a sparse VAR model by considering the maximum likelihood estimation with the sparsity constraints as a biconvex problem in the sense that the optimization problem becomes convex when either the autoregressive coefficients or the inverse covariance matrix is fixed.

ADVANCED SURVEY ESTIMATION METHODS IN WEB AND MIXED-MODE SURVEYS

CO014 **Room Sala 1** **Chair: David Molina**

CO0214: **Estimation techniques for discrete response variables in dual frame surveys**
Presenter: **David Molina**, University of Granada, Spain
Co-authors: Antonio Arcos, Maria Giovanna Ranalli

In a dual frame survey, two different samples are drawn; one from each of the sampling frames that are available, and then the information collected is adequately combined to get estimates. Although each of these frames may be incomplete considered separately, it is assumed that, jointly, they cover the entire target population. Since the emergence of dual frame surveys a noticeable number of estimators have been formulated to estimate the population total or mean of a continuous variable and, currently, dual frame surveys are widely used by statistical agencies and private companies due to their proven benefits. Surveys in general, and dual frame surveys in particular, usually include questions in which the respondents have to select one in a series of options. Although customary dual frame estimators can be used, the estimates they provide are inconsistent since they do not add up to 1 through all the possible categories of the response. New techniques for appropriately addressing variables with discrete outcomes in dual frame surveys are presented. To check the efficiency of the proposed procedures some Monte Carlo experiments were carried out.

CO0240: **Improving the item sum technique using auxiliary information in complex surveys**
Presenter: **Pier Francesco Perri**, University of Calabria, Italy
Co-authors: Beatriz Cobo

To collect sensitive data, survey statisticians have developed many approaches and strategies to reduce the rate of nonresponse and social desirability response bias. In the last years, the item count technique has been largely employed as alternative indirect questioning survey mode for qualitative characteristics and some variants have been proposed to face with new needs and challenges. The item sum technique (IST) is a recent variant introduced to estimate the mean of a sensitive quantitative variable when sampled units are asked to confront themselves with a two-list of items containing a question on the sensitive character under study and a number of innocuous questions. To the best of our knowledge, very few theoretical and applied works exist in the field of the IST. We, therefore, intend to discuss some methodological advances in order to spread the technique in real surveys. In particular, we discuss, under a generic sampling design, the problem of how to improve the estimates of the sensitive mean when auxiliary information on the population under study is available and used at the design and estimation stages. An Horvitz-Thompson type estimator and a calibration estimator are proposed and their efficiency evaluated on the basis of a simulation study performed on real data from the "World Bank's Enterprise Surveys". It is shown that estimates obtained by supposing that data are collected by the IST are nearly equivalent to those obtained using the true data.

CO0313: **Privacy protection in surveys with RR techniques using nonparametric regression**
Presenter: **Ismael Sanchez-Borrego**, University of Granada, Spain
Co-authors: Maria del Mar Rueda

People do not often respond truthfully when asked personal or sensitive questions in a survey, like those involving stigmatizing characteristics like regular gambling, marijuana consumption, tax evasion, etc. We consider the problem of estimating the finite population total of a quantitative variable using the randomized response (RR) technique, that preserves the privacy of the respondents. We propose a model-assisted estimator based on nonparametric regression, which can handle discrete and continuous data and is valid for any sampling design. The proposed method is shown to share some theoretical properties with the mixed-data kernel-based smoother in the survey context. We have investigated the practical performance of the proposed method under different scenarios with some randomization devices and different sampling designs. The nonparametric estimator is effective in estimating the population total of a continuous variable in simulation experiments in both natural and artificial populations.

SPATIAL STATISTICS AND DYNAMIC NETWORKS

CO055 **Room Sala 2** **Chair: Papa Ousmane Cisse**

CO0301: **The seasonal fractionally integrated separable spatial autoregressive model and its properties**
Presenter: **Papa Ousmane Cisse**, Gaston Berger, Senegal
Co-authors: Abdou Ka Diongue, Dominique Guegan

A new model is introduced called Fractionally Integrated Separable Spatial Autoregressive processes with Seasonality and denoted Seasonal FISSAR for two-dimensional spatial data. We focus on the class of separable spatial models whose correlation structure can be expressed as a product of correlations. The studies of spatial data have often shown presence of long-range correlation structures. To deal with this specific feature some authors had extended the long memory concept from times series to the spatial context. Thus it seems natural to incorporate seasonal patterns into the spatial model as soon as we work with data collected during many periods. This new modelling will be able to take into account periodic and cyclical behaviours presented in a lot of applications including the modelling of temperatures when the data are collected during different seasons at different locations. We investigate the properties of this new model providing stationary conditions, some explicit expressions form of the autocovariance function and the spectral density function. We establish the asymptotic behaviour of the spectral density function near the seasonal frequencies and perform some simulations. Some methods for estimating the parameters of the Seasonal FISSAR model are also discussed.

CO0339: **Cross-sectional analysis through rank-based dynamic portfolios**
Presenter: **Ludovic Cales**, Joint Research Centre - European Commission, Italy
Co-authors: Dominique Guegan, Monica Billio, Ludovic Cales

The aim is to study the cross-sectional effects present in the market using a new framework based on graph theory. Within this framework, we represent the evolution of a dynamic portfolio, i.e. a portfolio whose weights vary over time, as a rank-based multivariate model where the predictive ability of each cross-sectional factor is described by a variable. Practically, this modeling permits us to measure the marginal and joint effects of different cross-section factors on a given dynamic portfolio. Associated to a regime switching model, we are able to identify phases during which the cross-sectional effects are present in the market.

CO0343: **Time varying graphs**
Presenter: **Matteo Iacopini**, Ca Foscari University of Venice, Italy
Co-authors: Dominique Guegan

A method is proposed to study the dynamics of the dependence relation among a set of random variables, whose conditional independence relationships can be described by means of a graph. The procedure consists in five steps and exploits Vine Copula for representing the joint density of the set of RVs at each time, then a suitable invertible transformation is applied to map this distribution function into a tractable new space. The dynamics is introduced by defining and estimating an autoregressive process in this space, with the aim of making a one step ahead forecast. Finally, the invertibility of the transformation allows to obtain the predicted copula. The resulting distribution can be directly studied (global approach) by means of the relevant statistics; in addition, we would propose a way to infer the corresponding change of the graphical structure (specific approach). The method we propose is almost nonparametric, proving high flexibility in modeling the dependence relations among RVs, while a

simple AR process is dened for modeling time variation, which facilitate the interpretation without constraining.

METHODS AND COMPUTATIONS IN STATISTICS

CC068 **Room Sala 5** **Chair: Karel Hron**

CC0162: **Nonparametric hypothesis testing for isotonic survival models with clustering**
Presenter: **John Eustaquio**, University of the Philppines - Diliman, Philippines

Nonparametric hypothesis testing procedures based on the bootstrap were developed in testing for constant clustering effect in a survival model that incorporates the clustering effect into the Cox Proportional Hazards model. In a clustered survival model, bootstrap estimators of the cluster-specific parameters are consistent. Simulation studies indicate that the procedure is correctly-sized and powerful in a reasonably wide range of data. The test procedure for constant cluster effect over time is also robust to model misspecification. In survival data characterized with large number of clusters, the test is powerful even if the data is highly heterogenous and/or there is misspecification error.

CC0251: **Approximating the Rao's distance between negative binomial distributions: Application to counts of marine organisms**
Presenter: **Claude Mante**, Aix-Marseille University, France
Co-authors: Saikou Kide

While the negative binomial distribution is widely used to model catches of animals, it is noteworthy that the parametric approach is ill-suited from an exploratory point of view. Indeed, the "visual" distance between parameters of several distributions is misleading, since on the one hand it depends on the chosen parametrization and on the other hand these parameters are not commensurable (i.e. they measure quite different characteristics). Consequently, we settle the topic of comparing abundance distributions in a well-suited framework: the Riemannian manifold of negative binomial distributions, equipped with the Fisher-Rao metrics. It is then possible to compute an intrinsic distance between species. We focus on computational issues encountered in computing this distance between marine species.

CC0475: **Classification based on dissimilarities**
Presenter: **Beibei Yuan**, Leiden University, Netherlands
Co-authors: Willem Heiser, Mark De Rooij

The δ-machine is introduced. This is a statistical learning tool for classification based on dissimilarities or distances, δ, between inputs. The first step is to compute Euclidean distances between objects based on the predictor variables. Thereafter, we define varuous functions of the distances that produce (dis)similarity kernels. We distinguish four functions: the identity function, the squared function, the exponential decay function, and the Gaussian decay function. The (dis)similarity-kernels take the role as predictors in classification techniques. Classification decisions are based on the dissimilarity of objects to the selected exemplars or prototypes. This leads to nonlinear classification boundaries in the original predictor space. In a simulation study we compare the different dissimilarity-based logistic regressions using three types of artificial data. One with linear classification boundaries and two with nonlinear boundaries. Furthermore, we investigate the effect of noise predictor variables, the effect of sample size, and the effect of the number of predictors. The simulation study shows that overall three kernels perform very well (all but the squared Euclidean), and that these kernels are very flexible in the type of data they can handle.

STATISTICS FOR SCIENTIFIC PERFORMANCE EVALUATION

CG038 **Room Sala 3** **Chair: Ana Belen Ramos-Guajardo**

CC0520: **An assessment of scientific research performance for ranking countries from EU**
Presenter: **Florentin Serban**, Bucharest University of Economic Studies, Romania
Co-authors: Anca-Teodora Serban-Oprescu

The design of a balanced development strategy for a certain activity of an organization entails creating rankings of various entities relevant for that activity. The concept of entropy is widely used in decision-making problems as a useful instrument to assess the amount and effect of information provided by certain criteria used to construct a composite indicator. We propose a practical approach to evaluate academic performance, which can be applied when the values of several indicators are available for every EU country under observation. The method is based on the construction of a composite indicator, defined as the weighted sum of the indicators considered in the study. The weights of the indicators stand as a measure of importance for each criterion involved in defining the composite indicator. We present an application which aims to analyze several countries within EU in terms of their scientific research output. Numerical results for the considered EU countries are presented.

CC0534: **An integrative framework for evaluating scientific research**
Presenter: **Anca-Teodora Serban-Oprescu**, Bucharest University of Economic Studies, Romania
Co-authors: Luiza Badin, Silvia Dedu, Florentin Serban

Latest trends in evaluating performance in Research and Development show that advanced methodologies in econometrics and operations research have proved extremely useful and relevant in analyzing efficiency and productivity of various activity types, including scientific research. The aim is to review the latest conceptual approaches, methodologies and research tools from a complex inter and trans-disciplinary perspective. Advocating for the assessment of scientific research in an integrative manner, subscribed to the logic of a society based on knowledge and innovation, we propose a general framework for analyzing scientific research performance based on flexible nonparametric models which include exogenous factors with essential role in research evaluation. Taking into account cross-disciplinary factors considered so far in separate contexts, the analysis may lead towards a broader and multifaceted appraisal of scientific research.

CC0554: **Exponentiated-type distribution families with applications to scientific performance evaluation**
Presenter: **Silvia Dedu**, Bucharest University of Economic Studies, Romania
Co-authors: Vasile Preda

New exponentiated-type families of distributions are proposed and their properties are studied. The most relevant probabilistic descriptive measures are obtained, including moments, quantiles and concentration measures. Also, the distribution of order statistics is derived and several estimation methods are performed, based on moments, quantiles, least squares and maximum likelihood estimation. The relevance and effectiveness of the theoretical results are illustrated by solving several problems which consist in modeling and analyzing informetric distributions. The applications on real data sets containing impact factors from several relevant scientific fields prove that the use of these distribution families allows a higher degree of flexibility. Computational results are provided in order to highlight the advantages of our models.

BAYESIAN METHODS

CC065 **Room Sala 4** **Chair: Simon Wilson**

CC0420: **Inference in nonlinear systems with unscented Kalman filters**
Presenter: **Diana Giurghita**, University of Glasgow, United Kingdom
Co-authors: Dirk Husmeier

An increasing number of scientific disciplines, most notably the life sciences and health care, have become more quantitative, describing complex systems with coupled nonlinear differential equations. While powerful algorithms for numerical simulations from these systems have been developed, statistical inference of the system parameters is still a challenging problem. A promising approach is based on the unscented Kalman

filter (UKF), which has seen a variety of recent applications, from soft tissue mechanics to chemical kinetics. We investigate the dependence of the accuracy of parameter estimation on the initialisation. Based on three toy systems that capture typical features of real-world complex systems - limit cycles, chaotic attractors and intrinsic stochasticity - we carry out repeated simulations on a large range of independent data instantiations. Our study allows a quantification of the accuracy of inference, measured in terms of two alternative distance measures in function and parameter space, in dependence on the initial deviation from the ground truth.

CC0357: **Sequential importance sampling for online Bayesian changepoint detection**
Presenter: **Lida Mavrogonatou**, University of Glasgow, United Kingdom
Co-authors: Vladislav Vyshemirsky

Online detection of abrupt changes in the parameters of a generative model for a time series is useful when modelling data in areas of application such as finance, robotics, and biometrics. We present an algorithm based on Sequential Importance Sampling which allows this problem to be solved in an online setting without relying on conjugate priors. Our results are exact and unbiased as we avoid using posterior approximations, and only rely on Monte Carlo integration when computing predictive probabilities. We apply the proposed algorithm to three example data sets. In two of the examples we compare our results to previously published analyses which used conjugate priors. In the third example we demonstrate an application where conjugate priors are not available. Avoiding conjugate priors allows a wider range of models to be considered with Bayesian changepoint detection, and additionally allows the use of arbitrary informative priors to quantify the uncertainty more flexibly.

CC0569: **Discrete Bayesian DAG models with a restricted set of directions**
Presenter: **Jacek Wesolowski**, Warsaw University of Technology, Poland

First, we develop a new family of conjugate prior distributions for the cell probabilities of discrete graphical models Markov with respect to a set P of moral directed acyclic graphs (DAGs) with skeleton a given decomposable graph G. This family, called P-Dirichlet, is a generalization of the hyper Dirichlet: it keeps the directed strong hyper Markov property for every DAG in P but increases the flexibility in the choice of its parameters. Second, we prove a characterization of the P-Dirichlet, which yields, as a corollary, a characterization of the hyper Dirichlet as well as the classical Dirichlet law. Like the characterization of the classical Dirichlet, our result is based on local and global independence of the probability parameters. We use also separability property explicitly defined but implicitly used by Geiger and Heckerman through their choice of two particular DAGs. Another advantage of our approach is that we need not make the assumption of the existence of a positive smooth density function. We use the method of moments for our proofs.

POSTER SESSION II

CP107 **Room Ground Hall** **Chair: Cristian Gatu**

CP0413: **A multimomental ARMA model: Initial formulation and a case study**
Presenter: **Thomas Michael Bartlett**, University of Campinas, Brazil
Co-authors: Levy Boccato

Recently, nonnormal distribution functions have been developed to model more precisely and richly the behavior of time series of data. We aim at developing a times series model that describes, by mixing Gaussian distributions, the evolution of each statistical moment up to the third order - mean, variance and skewness. To each instant of time and to each moment an ARMA law of evolution is applied and the estimation of the model is done by optimizing a quasi likelihood function that depends on the ARMA coefficients of the three moments. Thus, the method of moments for Gaussian mixtures is used to obtain a probability density function which has the desired instantaneous moments. Employing Newton Method and Nelder-Mead optimization procedures to estimate the model, an empirical analysis is done studying the model's performance and consistency using a synthetic time series.

CP0160: **Relevance of combining ARFIMA and artificial intelligence-based models: An empirical evidence from Scandinavian market**
Presenter: **Najeh Chaabane**, Sousse-ISFF, Tunisia

Electricity prices involve many features comparably with those in financial markets. In the framework of competitive electricity market, forecasting such prices has become a real challenge for all market participants. Therefore, the choice of the forecasting model has become even more important. ARFIMA and Artificial Intelligence-Based Models are proposed. Such models can deal simultaneously with long memory behavior and non-linearity existing in electricity spot prices. Data from Scandinavian market are used to synthesize and examine the strength of the proposed models.

CP0502: **Multiple use confidence intervals for the statistical calibration problem**
Presenter: **Martina Chvostekova**, Institute of Measurement Science of Slovak Academy of Sciences, Slovakia

The statistical calibration problem consists of constructing the interval estimates for future unobserved values of an explanatory variable, say x, corresponding to possibly infinitely many future observations of a response variable, say y, based on n pairs of values (x_i, y_i), $i = 1, 2, \ldots, n$. In the basic setting of the model it is assumed a linear dependence between the variables, a polynomial regression is probably the most frequently used model in industrial applications, and the measurements errors of the observation variable are independent normally distributed variables with zero mean and a common unknown variance. A marginal property of the multiple use confidence intervals is that at least γ proportion of them contains the corresponding true value of the explanatory variable with confidence $1-\alpha$. One standard way to construct the multiple use confidence intervals is to invert $(\gamma, 1-\alpha)$ simultaneous tolerance intervals for a linear regression, but such multiple use confidence intervals are conservative. It is proposed a procedure for determining the exact multiple use confidence intervals under assuming a distribution of the explanatory variable. The computation of the suggested multiple use confidence intervals is fast and they are uniformly narrower than previously published.

CP0521: **Dilemma: Distributed learning with Markov Chain Monte Carlo algorithms**
Presenter: **Ali Zaidi**, Microsoft, United States

Many scalable Monte Carlo algorithms conduct local updates across batched datasets to construct reduced variance estimators that are inherently biased. By reweighting the samples processed by each node in a distributed clustering environment, it is possible to reduce the overall bias of the algorithm while attaining computationally efficient and low variance estimators. However, theoretical guarantees on the convergence of the algorithm (or the infinitesimal generator of the algorithm) to the true target distribution are no longer valid due to the asymptotic bias induced by such distributed computations. Stein's method for bounding divergence of measures is described to compute the discrepancy between the target distribution and sampled distribution. It further supplements this methodology with the Malliavin calculus for estimating functionals of Gaussian processes, in order to utilize the discrepancy measure to produce automatically tuned Meteropolis proposals to optimize exploration/exploitation schemes for complex target distributions. The aim is to summarize these results, to describe how to use the accompanying R package, and to provide examples for the overall distributed learning framework.

CP0539: **Capturing diagnostic error in geostatistical models of Malaria survey data**
Presenter: **Eleni Verykouki**, Swiss Tropical and Public Health Institute, Switzerland
Co-authors: Andres Cardona Gavaldon, Penelope Vounatsou

Malaria is a blood disease that is caused by parasites that are transmitted to humans via the Anopheles mosquito. It is usually found in tropical and subtropical climates where the parasites that cause it live, and it is most prevalent in the African region. Children under five years of age are one of most vulnerable groups affected by malaria. Early and accurate diagnosis of malaria is essential for effective disease management and malaria

surveillance. There are two main diagnostic tools to test for malaria, Microscopic diagnosis and Rapid Diagnostic Test (RDT). The aim concerns the estimation of the prevalence of malaria in children under five in Africa. Data consist of environmental and socio economic parameters. Data of microscopy tests and RDT are also used to measure for diagnostic error in the case where neither tests can be considered as a gold standard. Bayesian inference and spatial statistical methodology are employed to obtain high spatial resolution maps of malaria transmission risk in children under five in African countries.

CP0542: **Classification of individual disability progression trajectories of multiple sclerosis patients**
Presenter: **Ceren Tozlu**, University of Lyon, France
Co-authors: Gabriel Kocevar, Francoise Durand-Dubief, Sandra Vukusic, Dominique Sappey-Marinier, Delphine Maucort-Boulch

Multiple sclerosis is a demyelinating, inflammatory, chronic disease of the central nervous system. MS occurs in 4 subtypes such as the clinically isolated syndrome, remitting-relapsing, secondary progressive and primary-progressive. The course of the disease is very different between patients. Therefore, the major challenge of to-days neurologist is to classify patients. The diffusion tensor imaging is an effective mean for the quantification of demyelination and axonal loss measured with fractional anisotropy, axonal, radial and mean diffusivity markers. 80 MS patients divided in 4 subtypes were included in the study. The patients were followed up with standardized clinical and MRI examination every six months during the first 3 years and every year during the last 2 years. The imaging data is obtained in 5 regions of Corpus Callosum. The kml and kml3d packages are proposed to cluster and to choose the best number of clusters for respectively the clinical and imaging trajectories. The imaging trajectories are classified for all couple of markers at each CC region. The best number of clusters is obtained as 3 for clinical data as well as for DTI data which is generated with fractional anisotropy and mean diffusivity markers at 4 regions of CC. Although 4 subtypes are predefined, 3 clusters of clinical data may be explained by the evolution of CIS patients afterwards as RR. The packages provided promising classifications which correspond well to the clinical classifications.

CP0519: **Estimating the effectiveness of AdWords geo-targeting**
Presenter: **Iman Al-Hasani**, Durham University, United Kingdom

AdWords is an online advertising service, allows advertisers to place advertising. AdWords allows advertisers to track the performance of their ads by receiving detailed data such as website visits, number of clicks, sales and the revenue. These metrics help to measure the effectiveness of the advertising. In practice, however the actual effects of the ads campaign is difficult to quantify due to the noise in users behaviour. A comparison geo experiment approach has been used to measure the impact of the advertising . In this experiment a region of interest is partitioned into a set of geographical areas, called geos. These geos are then assigned randomly to serve the ads. The aim is to provide extended statistical methodology for geo experiment. It is expected to suggest improvements in generation the geos, construct a proper design of how to allocate these geo to treatment group and estimate the size of the effect of targeting within each geo. Adwords geo-targeting are linked to governmental areas using the shortest distance over the earth's surface. A simulation structure has been constructed to simulate the online searches and purchases occured at each target geo. Logistic regression is used to estimate the probability of search and conditionally the probability of purchase. The size of the effect of the ad campaign is computed for different campaign design structures. The estimation are computed based on covariates attributable to individuals, regions and ad campaign.

CP0365: **Application of ensemble methods for detection of changepoints in generalized linear models**
Presenter: **Asanao Shimokawa**, Tokyo University of Science, Japan
Co-authors: Takuma Kurosawa, Etsuo Miyaoka

The focus is on the estimation of the number of changepoints and detection of their locations under a generalized linear model. If the number of changepoints is small and the size of data is not so large, there are several proposed methods for detecting them, like hierarchical binary splitting algorithm or dynamic programming algorithm. However, if the size of data is too large to search all possibility of change points and/or if it is necessary to estimate the number of change points, these algorithms are difficult to run. To address these problems, we consider to use ensemble methods for estimating the number of change points and detecting those locations at same time. The performance of these methods are examined by simulation studies. Additionally, we show the results of applying the methods to actual data.

CP0477: **Bivariate regional frequency analysis of extreme precipitation events in the Czech Republic**
Presenter: **Tereza Simkova**, Technical University of Liberec, Czech Republic
Co-authors: Jan Picek, Jan Kysely

Bivariate regional frequency analysis (RFA) has been applied to estimate quantile curves for simultaneous non-exceedance probabilities $p = 0.9, 0.95, 0.99, 0.995, 0.999$ and joint AND/OR return values corresponding to 10-, 20-, 50- and 100-year return periods of extreme precipitation events in the Czech Republic. Maximum annual 1- and 5-day precipitation totals measured from 1961 to 2007 at 210 stations were used as the input dataset. When determining the bivariate distribution function of the 1- and 5-day maximum annual precipitation amounts, standard approaches were followed involving the goodness-of-fit test based on the Cramér-von Mises statistic, Akaike Information Criterion and estimation of the upper-tail dependence coefficient and considering a wide copula test space. The Hüsler-Reiss copula family has been identified as the most suitable for modelling the link between variables. Only in the northernmost part of the area the Gumbel and Joe copula families have been preferred. It has been found out the homogeneity condition (i.e. that sites in a region have the probability distribution identical apart from a site-specific scale factor) is satisfied for every of the six proposed regions and considered bivariate distribution functions, and therefore the index-flood based bivariate RFA was preferred to use. The resulting models, bivariate quantiles and joint return values provide valuable information and can be useful for hydraulic engineering design and planning.

CP0387: **Relationship between dietary habits and dementia status among Japanese suburban community-dwelling elderly**
Presenter: **Chisako Yamamoto**, Shonan University of Medical Sciences, Japan

The aim is to clarify the relationship between dietary habits and dementia status in the community-dwelling elderly aged 65 years and older. Self-administered questionnaires were sent to 2,069 elderly in March 2004 and 1,538 were returned by mail (response rate 74.3%). The subjects answered frequency of dietary intakes of meat/poultry, soy products, eggs, oily fish, dairy products, fruits, vegetables, fats and alcohol. We classified analysis subjects into three groups according to intellectual activity scores: people with dementia (PWD), people with probable dementia (PPD) and cognitively intact people (CIP). The chi-square, Kruskal-Wallis, Mann-Whitney U and Bonferronis multiple comparison tests were performed. Significance was set at 0.05. The Kruskal-Wallis and Mann-Whitney U tests revealed significant differences in women between CIP and PPD in all dietary items and between CIP and PWD in soy products, oily fish and vegetables. Men showed significant differences between CIP and PPD in fruits and alcohol and between CIP and PWD in fats and alcohol. The conclusion is that dietary items above including moderate amount of alcohol are recommended to prevent or delay the onset of dementia by Alzheimers Society. The results suggest that CIP womens health behavior was better than mens, supporting the recommended foods as healthy diet.

CP0525: **Generation of monthly precipitation series and the performance of homogeneity tests**
Presenter: **Elif Akca**, Middle East Technical University, Turkey
Co-authors: Ceylan Yozgatligil, Ceyda Yazici

Climate studies have gained importance due to the significant effect of climate change. The extreme meteorological events can cause floods, droughts, sudden change in the temperature or change in the climate trends. Since they have an important effect on human beings and the environment, these meteorological variables should be predicted and some precautions should be taken if possible. In order to conduct any kind of

statistical analysis, the nonclimatic effects should be determined and corrected or removed. Otherwise, any kind of meteorological effect should be kept to reflect the change in the variable. If there is any nonclimatic effect in the variable, the series is defined as inhomogeneous. The most important climatic variables are temperature and precipitation. The precipitation variable cannot be modeled efficiently due to its nonsymmetric shape and high variability. First, monthly total precipitation amount is simulated based on real station records. Then, in each simulation, artificial breakpoints are created and the homogeneity analysis are conducted to detect the inhomogeneity and the results are compared to decide on the best performed homogeneity test for precipitation.

CP0436: **Comparison of group means using the modified generalized F-test in the presence of outliers under heteroskedasticity**
Presenter: **Mustafa Cavus**, Anadolu University, Turkey
Co-authors: Berna Yazici, Ahmet Sezer

There are many powerful methods in the literature to test the equality of group means under heteroskedasticity. These methods lose their power in the presence of outliers. Handling this problem, some robust modifications can be applied to these methods. We show the power of the modified generalized F-Test to resist to outlier(s). The maximum likelihood estimators of location and scale parameters are replaced with Huber's M-estimation in the generalized F-Test to handle outlier(s). The efficiency of this modification is pointed out by a real data application. The data concerns the 2015 Annual Export Amounts of cities in Turkey, obtained from the Turkish Statistical Institute Database. Although the results of the F-test indicate equality of group means in the presence of outlier(s) under heteroskedasticity, the modified Generalized F-test revealed the difference between the group means. These results are supported by post-hoc comparison tests. The modified generalized F-test should be preferred over the other methods under heteroskedasticity with outlier(s).

CP0538: **Improvement of the methods of providing meteorological data for energy management systems**
Presenter: **Takamitsu Funayama**, Tokai University, Japan
Co-authors: Takeshi Watanabe, Hideaki Takenaka, Hideki Kimura, Kota Fukuda, Takashi Nakajima, Yoshiro Yamamoto

To grasp solar radiation for each area is essential for Energy Management Systems (EMS), including solar power. Solar radiation using satellite data from Himawari-7 was provided in the past in binary data format. The data can be converted to several formats using tools for Mathlab, Fortran and so on. Most of the energy demand researchers want to get a specific area data. To use such data providing methods, the researcher needs to download a lot of unnecessary data and has to program. Since 2015 August, we have got high-frequency and high-resolution data from Himawari-8, then solar radiation data becomes large size and is providing in intervals. We developed a web-based data provider system which has a facility to select a specific city or area. We introduce an example of providing meteorological data to a moving vehicle to help the solar car team from our university to give them meteorological information.

CP0471: **Comparison of two regression models using the generalized $p-$value method**
Presenter: **Seray Mankir Kahvecioglu**, Anadolu University, Turkey
Co-authors: Berna Yazici

In regression analysis, in case of comparing two regression models and coefficients where the distribution of variables is not known, generalized $p-$values may be used. The generalized $p-$value are an extended version of the classical $p-$value which provides only approximate solutions. Using approximate methods, generalized $p-$value, has better performance with small samples. The generalized $p-$values, which may be used alternatively when different assumptions, are not fulfilled, are investigated theoretically; a suitable simulation program is written and an application in regression analysis is given.

CP0366: **Clustering for multivariate functional data**
Presenter: **Pai-Ling Li**, Tamkang University, Taiwan
Co-authors: Ling-Cheng Kuo

A novel multivariate k-centers functional clustering algorithm for the multivariate functional data is proposed. We assume that clusters can be defined via the functional principal components subspace projection for each variable. A newly observed subject with multivariate functions is classified into a best-predicted cluster by minimizing a weighted distance measure, which is a weighted sum of discrepancies in observed functions and their corresponding projections onto the subspaces for all variables, among all the clusters. The weight of each variable represents the importance of a variable to the cluster information and is determined by the within-variable variation or the between-variable correlations. The proposed method can take the means and modes of variation differentials among groups of each variable into account simultaneously. In addition, the weight of the proposed algorithm is flexible and can be chosen by the objective of clustering. Numerical performance of the proposed method is examined by simulation studies, with an application to a data example.

CP0453: **Geographically weighted quantile regression for count data**
Presenter: **Vivian Yi-Ju Chen**, Tamkang Universiy, Taiwan

Past decade have witnesses an explosion both in the applications of Geographically Weighted Regression (GWR) and of Quantile Regression (QR). While these two techniques have become the commonplaces in many disciplines, they have never been integrated until an analytic framework called geographically weighted quantile regression (GWQR) has been proposed recently. The current structure of GWQR is however restricted to the analysis of continuous dependent variables. Discrete count data are observed in many fields such as health (disease counts), transportation (accidents), and finance (number of bankruptcy). When it comes to model such type of outcome, GWQR is inappropriate and provides insufficient information of the data. We aim to address the gap by extending GWQR for continuous dependent variables to the generalized GWQR framework for count variables. We first formulate the modeling specification, and then develop bootstrap methods to conducting the inference of model parameters. Finally, the proposed model is applied to a dataset of dengue fever in Taiwan as an empirical illustration.

ROBUSTNESS FOR HIGH-DIMENSIONAL DATA

CI077 **Room Sala Camara** **Chair: Stefan Van Aelst**

CI0192: **Detecting anomalous data cells**

Presenter: **Peter Rousseeuw**, KU Leuven, Belgium

Co-authors: Wannes Van den Bossche

A multivariate dataset consists of n cases in d dimensions, and is often stored in an n by d data matrix. It is well-known that real data may contain outliers. Depending on the situation, outliers may be (a) undesirable errors which can adversely affect the data analysis, or (b) valuable nuggets of unexpected information. In statistics and data analysis the word outlier usually refers to a row of the data matrix, and the methods to detect such outliers only work when at least half the rows are clean. But often many rows have a few contaminated cell values, which may not be visible by looking at each variable (column) separately. A method to detect anomalous data cells is proposed, which takes the correlations between the variables into account. It has no restriction on the number of clean rows, and can deal with high dimensions. Other advantages are that it provides estimates of the 'expected' values of the outlying cells, while imputing missing values at the same time. The method is illustrated on several real data sets, where it uncovers more structure than found by purely columnwise methods or purely rowwise methods. It can also serve as an initial step for estimating multivariate location and scatter matrices.

CI0294: **On robust and nonparametric change-point detection in multiple time series**

Presenter: **Roland Fried**, TU Dortmund University, Germany

Co-authors: Alexander Duerre, Herold Dehling, Daniel Vogel, Martin Wendler

A basic issue in the analysis of time series data is the question of stability of the data generating process. Especially in multivariate time series, there are many different aspects which can change over time. Methods for the detection of abrupt changes at unknown time points have been developed during several decades and offer some power also against monotonic drifts. Most of the existing work is based on linear statistics and/or for the situation of a single time series. Starting from previous work on robust and nonparametric change-point detection in univariate time series based on the usage of U-statistics and U-quantiles, a general class of tests for the detection of changes of possibly several types in multivariate time series is developed and investigated in different scenarios concerning the type of time series model and the type of structural change.

CI0333: **Robust principal components for high-dimensional data**

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

Co-authors: Holger Cevallos Valdiviezo, Matias Salibian-Barrera

Classical (functional) principal component analysis can be written as a least squares optimization and thus can be highly influenced by outliers. To reduce the influence of atypical measurements in the data, we propose two methods based on trimming: a multivariate least trimmed squares (LTS) estimator and a componentwise variant. The multivariate LTS minimizes the least squares criterion over subsets of observations. The componentwise version minimizes the sum of univariate LTS scale estimators in each of the components separately. The methods can directly be applied to high-dimensional multivariate data. Instead of LTS scales other robust scales such as S-scales can be considered as well. The methods can also be applied to functional data. In the case that the curves are irregularly spaced, a smoothing step can be applied to represent the curves in a high-dimensional space. The resulting solution is then mapped back onto the functional space. Outliers can be identified by examining their orthogonal distance from the subspace.

STATISTICAL EVALUATION OF MEDICAL DIAGNOSTIC TESTS

CO002 **Room Sala 2** **Chair: Maria del Carmen Pardo**

CO0199: **An index to analyze a portion of the ROC curve**

Presenter: **Alba Franco-Pereira**, Universidad de Vigo, Spain

Co-authors: Maria del Carmen Pardo

In clinical epidemiology, the partial Area Under the Curve (pAUC) at a high specificity threshold has been proposed as an indicator of test accuracy when the ROC curve is asymmetric. For example, acceptable specificities are high for early cancer detection tests. A lower specificity for a large population leads to many more falsely classified non-diseased subjects who may have to undergo a more invasive test subsequently. It is thus desired to compare screening markers at a higher range of specificities. The pAUC, which summarizes part of the ROC curve in the range of desired specificities, use to be the alternative to AUC. Several methods have been developed as alternative to pAUC but most of them assume the data have an underlying normal distribution. We propose a nonparametric partial summary index and its properties are explored by simulations. Finally, the new index is illustrated with a real data set.

CO0196: **Optimal cutpoints for classification in medical diagnostic tests**

Presenter: **Monica Lopez Raton**, Conselleria de Educacion- Xunta de Galicia, Spain

Co-authors: Carmen Cadarso Suarez, Elisa Maria Molanes Lopez

Continuous diagnostic tests (biomarkers or risk markers) are often used to discriminate between healthy and diseased populations. For their clinical application, the key aspect is how to select an appropriate cutpoint or discrimination value c that defines positive and negative test results. In general, individuals with a test value smaller than c are classified as healthy and otherwise as diseased. In the literature, different optimality criteria there exist to select c. We consider an interesting cost-based generalization of the Symmetry point (the optimal c that maximizes simultaneously both types of correct classifications) incorporating the misclassification costs, and we propose confidence intervals for this optimal cutpoint and its sensitivity and specificity measures using two approaches: a parametric approach based on the Generalized Pivotal Quantity (GPQ) under normality and a nonparametric approach based on the Empirical Likelihood (EL). In addition, we develop two R packages, OptimalCutpoints and GsymPoint, to facilitate clinicians selecting optimal cutpoints in their daily practice. A new classification rule is also proposed by logistic generalized additive regression models (GAMs), that provides an improved discriminatory capacity where traditional Receiver Operating Characteristic (ROC) analysis is not valid, being necessary more than one optimal cutpoint on which to base the classification.

CO0262: **Confidence intervals for differences between volumes under ROC surfaces and generalized Youden indices**

Presenter: **Christos T Nakas**, University of Bern, Switzerland

Co-authors: Benjamin Reiser, Lili Tian, JingJing Yin

Receiver Operating Characteristic (ROC) surfaces are used in three-class diagnostic testing especially in settings were an ordering between classes exists. The ROC surface is a 3D plot of the true-class fractions for the classes under study over all possible threshold values (c_1 and c_2, with c_1 less than c_2) of a diagnostic marker. To evaluate the discriminatory ability of a marker we summarize the information of an ROC surface into a single global index. The Volume Under the ROC surface (VUS) and the Generalized Youden Index (GYI) are two such indices. For the comparison of diagnostic markers the difference between VUSs or the difference between GYIs can be considered. We describe and compare parametric and nonparametric methods for inference based on the difference between VUSs and GYIs in a paired setting where all subjects under study undergo two diagnostic tests and interest lies in comparing these diagnostic tests. Parametric methods include the use of the asymptotic delta method under normality or after applying the Box-Cox transformation for data normalization, and the use of Generalized Pivotal Quantities. Different variants

of these methods are studied. Nonparametric methods involve empirical estimation and kernel smoothing. We compare methods through extensive simulations and discuss an application.

CO0289: **Identifying optimal biomarker combinations for treatment selection using data from randomized controlled trials**
Presenter: **Ying Huang**, Fred Hutchinson Cancer Research Center, United States

Biomarkers associated with treatment-effect heterogeneity can be used to make treatment recommendations that optimize individual clinical outcomes. Statistical methods are needed to generate marker-based treatment-selection rules that can effectively reduce the population burden due to disease and treatment. Compared to the standard approach of risk modeling to combine markers, a more robust approach is to directly minimize an unbiased estimate of total disease and treatment burden among a pre-specified class of rules. We frame this into a general problem of minimizing a weighted sum of 0-1 loss and propose a penalized minimization method based on the difference of convex function algorithm, using data from randomized trials. The corresponding estimator has a kernel property that allows flexible modeling of linear and nonlinear combinations of markers. We further expand the method with a L1-penalty to allow for feature selection and develop an algorithm based on the coordinate descent method. We compare the proposed methods with existing methods for optimizing treatment regimens such as the logistic regression, the weighted logistic regression, and the weighted support vector machine. Performances of different weight functions are also investigated. We illustrate the application of the proposed methods in host-genetics data from an HIV vaccine trial.

RECENT DEVELOPMENTS IN LATENT CLASS ANALYSIS AND ITS APPLICATIONS

CO061 **Room Sala 5** **Chair: Mattis van den Bergh**

CO0269: **Latent class trees**
Presenter: **Mattis van den Bergh**, Tilburg University, Netherlands
Co-authors: Jeroen Vermunt

Researchers use latent class analysis to derive meaningful clusters from sets of categorical observed variables. However, especially when the number of classes required to obtain a good fit is large, interpretation of the latent classes in the selected model may not be straightforward. To overcome this problem, we propose an alternative way of performing a latent class analysis, which we refer to as latent class tree modelling. For this purpose, we use a recursive partitioning procedure similar to those used in divisive hierarchical cluster analysis; that is, classes are split until the model selection criterion indicates that the fit does no longer improve. The key advantage of the proposed latent class tree approach compared to the standard latent class analysis approach is that it gives a clear insight into how the latent classes are formed and how solutions with different numbers of classes are linked to one another. We also propose measures to adjust the tree in certain conditions. The new approach is illustrated by the analysis of real data and simulation studies.

CO0284: **Goodness-of-fit in multilevel latent class analysis**
Presenter: **Erwin Nagelkerke**, Tilburg University, Netherlands
Co-authors: Daniel Oberski, Jeroen Vermunt

In the context of latent class models that deal with nested data, the goodness-of-fit depends on multiple aspects, amongst which several local independence assumptions. However, due to a lack of local fit statistics, the issues related to model fit can only be inspected jointly through global fit statistics. Given the number of possible model specifications, and potential areas for model misfit, the reliance on global fit poses several problems. Most notably, in case of a badly fitting model, there is no indication of the source of the misfit, which can prove valuable theoretical information. This is especially true for the fit of more complex models such as the multilevel model latent class model and the time dependencies in hidden Markov models. Vice versa, in case of an overall adequate fit, particular assumption violations may go unnoticed. For example, a multilevel latent class model may well be suited for the large majority of observed groups, leading to an overall good fit of the model, but could fail to model the dependencies of several extreme groups. New fit statistics are proposed to improve the understanding of the model, allow individual testing of the local independence assumptions, and inspect the fit of the model locally to pinpoint misfit and provide additional substantive insight.

CO0268: **New resampling methods applied to latent class model fit assessment**
Presenter: **Geert van Kollenburg**, Tilburg University, Netherlands
Co-authors: Joris Mulder, Jeroen Vermunt

The assessment of model fit is an important part of statistical analysis. The researchers interest may lie with specific aspects of a model, or in the global fit. Asymptotic $p-$values are not available for every conceivable statistic and even when they are available they may not be valid when sample sizes are not very large. To get more reliable $p-$values, researchers may resort to resampling methods. Some of these methods are time consuming, while others may provide $p-$values which are not uniform under the null-hypothesis. The most common resampling methods to test Latent Class model fit will be illustrated. A recently proposed calibration, the posterior predictive $p-$value, will be discussed. Finally a very fast resampling scheme is proposed where the statistics are based on data only, which requires that each model of interest is estimated only once.

CO0224: **Handling missing data with multiple imputation using LC models to investigate predictors of HPV infection**
Presenter: **Roberto Di Mari**, University of Rome Tor Vergata, Italy
Co-authors: Jlenia Caccetta, Maura Mezzetti

A statistical method is proposed to investigate whether various risk factors can successfully predict HPV infection, despite missing data, using data on a cohort of 864 women, with information on HPV risk level and 67 explanatory variables. Each record has at least one item non-response, with a total of 27% of missing values. Dealing with missing data, a very common issue in medical applications, is possible through multiple imputation (MI): once the missing values are imputed, the analysis can be done with standard techniques, without extra effort to interpret results. Whereas saturated log-linear models for imputation are unfeasible if the number of covariates is too large, the proposal is to approximate the conditional distribution of the missing data given the observed data using latent class analysis (LCA). This makes sure that, given a sufficiently large number of mixture components, complex associations between the variables are captured by the imputation model; in addition, covariates are included to further improve its accuracy. To reflect the uncertainty about the model parameters, non-parametric bootstrap is implemented. This is the first time that MI with LCA is used in a medical application, and the results obtained are in line with pathophysiology and the literature on HPV.

JAPANESE CLASSIFICATION SOCIETY INVITED SESSION: STATISTICAL ANALYSIS FOR CATEGORICAL DATA

CO057 **Room Sala 1** **Chair: Tadashi Imaizumi**

CO0306: **A model based clustering for ordered categorical data**
Presenter: **Tadashi Imaizumi**, Tama University, Japan

The volumes of data matrix have been larger and larger in modern data analysis. And the clustering methods will be useful to find several homogenous group from data. However, as we also treat many categorical data such text data in a newspaper, data collected by many sensors, etc., we need to think about how to model for analyzing these categorical data. One approach will be to adopt the model-based clustering approach. The latent class models and methods will be one alternative as model based approach. The categorical variables will be represented a binary vector corresponding to a categorical value of a categorical variables. So, the data matrix with p categorical variables will be a high dimensional and sparse data matrix. Though these latent class methods will be useful, but, they do not fit for data with larger sample cases or many categorical variables. And we need to develop a new model and methods of the latent class analysis or the model-based clustering for these type of data matrix.

A model and method will be proposed for uncovered the classes of n samples and classes of p categorical variables in frame of two-way clustering.

CO0411: **Time series changes of the categorical data using the text data regarding radiation**
Presenter: **Takafumi Kubota**, Tama University, Japan
Co-authors: Hitoshi Fujimiya, Hiroyuki A Torii

How the classified class of the questions keywords included in the text-based is changing in terms of time is discussed. After the accident of TEPCOs Fukushima Daiichi Nuclear Power Station in March 2011, a lot of information about the keyword relating to the radiation has been taken up in the media, and a group of people who receives the media has been becoming uneasy and dissatisfied about the information. To solve this issue, the Japan Health Physics Society started a website of questions and answers which related radiation (radiation QA). Within this, the experts have been answering towards the questions that the people asked (mainly the people in metropolitan area and the Fukushima prefecture). The goal is to hold a comparative verification by checking against one another of the various happenings that occurred after the incident with the related keywords of radiation, and also the groups the keyword belongs to, of time alongside with how the transitions were made, using the text data which is open to public on the radiation QA.

CO0404: **Risk profiles for severe mental health difficulty: Classification and regression tree analysis**
Presenter: **Yoshitake Takebayashi**, Institute of Statistical Mathematics, Japan
Co-authors: Takafumi Kubota, Tsubaki Hiroe

Severe mental health difficulty is a leading cause of suicide. Its development has been associated with several risk factors. Comprehensive approaches must be used to examine the degree to which these factors co-act and interact to develop severe mental health difficulty risk profiles. The literature review reveals no report of a study exploring interactions among severe mental health condition factors in subsets of people with high suicidal risk (mental disorders, unemployment, and caregivers of relatives) in Japan. We aim to use a classification and regression tree (CART) approach to establish risk profiles and examine their performance for diagnostic accuracy. Data were obtained from the National Comprehensive Survey of Living Conditions. Outcome measures (K6) were categorized into low, moderate, and high, applying the recommended cut-off values. Socio-demographic status, financial status, and subjective stress were included as predictors in the CART model. CART analysis results indicate that subjective stress in daily life is the strongest predictor for severe mental health difficulties in the high-suicide-risk group. Additionally, results show that all high-suicide-risk group divided into several sub-groups that reflect interactions among predictors.

CO0418: **Visualization of cross tabulation by the association rules by using the correspondence analysis**
Presenter: **Yoshiro Yamamoto**, Tokai Univeristy, Japan
Co-authors: Sanetoshi Yamada

When comparing the response in the survey by gender and age, we make the cross-tabulation tables then visualize them by such like mosaic plot. For many answers to multiple-choice item, we want to find the item that the reaction of a particular layer (gender and age) is different from the others. Association rule analysis are suitable for this kind of analysis. By using the coordinates by correspondence analysis it is possible to plot the relationship between items and media layers. In this visualization, correspondence analysis and association rule analysis are complementary to each other. In addition, by showing the percentage of respondents each item and each layer, it becomes possible to understand the trend between items and layers. The visualization is constructed by using RStudio Shiny. It is possible to change the various parameters of the association rule analysis interactively.

FUNCTIONAL DATA ANALYSIS

CC066 **Room Sala 3** **Chair: Ian McKeague**

CC0369: **A consistent estimator of the smoothing operator in the functional Hodrick-Prescott filter**
Presenter: **Hiba Nassar**, Lund University, Sweden

A version of the functional Hodrick-Prescott filter for functional time series is presented. We show that the associated optimal smoothing operator preserves the noise-to-signal ratio structure. Moreover, as the main result, we propose a consistent estimator of this optimal smoothing operator.

CC0541: **Registration method for functional data based on shape invariant model with t distribution**
Presenter: **Mariko Takagishi**, Doshisha University, Japan
Co-authors: Hiroshi Yadohisa

Registration for functional data aims at aligning each curves when the given data is misaligned, i.e. the beginning time point of the change is different depending on subjects. Shape Invariant Model (SIM) is a one of registration methods based on nonlinear regression models. Though the most existing SIM assume a normal distribution for the amplitude and phase variation, when there exist outliers, the estimation is often badly affected by them. Therefore, we propose SIM for $t-$distributions. The advantages of using $t-$distributions are; first, the effect of the outliers on the estimation is reduced since the $t-$distribution gives a bounded weight function. Second, since the model assuming normal distributions for amplitude and phase variation is nested with the one with assuming $t-$distributions, the update formula can be easily derived and the comparison with the existing normal model is simple.

CC0215: **Functional regression analysis with compositional response**
Presenter: **Renata Talska**, Palacky University Olomouc, Czech Republic
Co-authors: Alessandra Menafoglio, Karel Hron, Eva Fiserova, Jitka Machalova

Regression analysis is a key statistical tool to model a linear relationship between a response variable and a set of covariates. In functional data analysis (FDA), methods to perform linear regression with functional response and scalar predictors have been widely discussed. More delicate appears the situation in which the response variable is represented as a probability density, since the L2 space (of square integrable functions), usually employed in FDA, does not account for the inherent constraints of densities. The aim is to introduce functional regression model with distributional response using the Bayes space approach, i.e. a geometric viewpoint that allows capturing the inherent features of distributional data. Indeed, densities primarily carry relative information, and the unit integral constraint represents just one of its possible equivalent representations. Accordingly, densities can be considered as elements of a Bayes Hilbert space, whose geometry is designed to precisely capture the specific properties of densities (e.g. scale invariance, relative scale). In order to apply functional regression tools for L2 data, particularly those based on B-spline representations, the centred logratio transformation - mapping the Bayes Hilbert space into L2 - is considered. The methodological developments are illustrated with a real-world example.

CC0473: **Partially and dependently observed functional data**
Presenter: **Stefan Rameseder**, Regensburg, Germany
Co-authors: Dominik Liebl

As in the case of missing data in uni- and multivariate data sets, without assuming the so-called "missing-at-random" condition it is generally impossible to consistently estimate, e.g. the mean-function, if functional data are only partially observed. By contrast, for functional data there is a chance to consistently estimate the mean-function even though the "missing-at-random" assumption is violated. By using a detour via the fundamental theorem of calculus, we propose a new estimator of the mean-function for partially observed functional data. While we theoretically compare bias and variance of our estimator with typical mean estimators, we additionally perform an extended two-part simulation study. On the

one hand, we investigate the applicability of our estimator via an identification procedure by sequential testing. On the other hand, we consider bias and variance of our estimator versus other estimators in different missing data scenarios. As an empirical motivation, we apply this procedure onto supply curves in a frequential multi-unit auction with pay-as-bid pricing mechanism and exogenous and predetermined demand. In this market design, typical trading strategies strongly depend on the preannounced demand for which our estimator in opposite to others delivers useful results for the whole mean curve.

ANALYSIS OF SPATIAL AND TEMPORAL DATA

CG009 **Room Sala 4** **Chair: Marc Genton**

CC0380: **Detection of space-time clusters for radiation data using spatial interpolation and scan statistics**
Presenter: **Fumio Ishioka**, Okayama University, Japan
Co-authors: Koji Kurihara

On March 11, 2011, a massive amount of radioactive material was released into the environment because of the Fukushima Daiichi Nuclear Power Station (NPS) accident. Surveys on the amount of radioactive materials are very important for assessing the state of the surrounding environment and planning future countermeasures. We attempt to detect a high-contaminant cluster accompanied by a time change for the area of evacuation in the Fukushima Prefecture from January 10-19, 2013. The data were air dose rate measured by monitoring post. As a priori analysis, we apply a spatial interpolation using ordinary kriging, because the observations obtained were very sparsely scattered and had extremely large dispersion and bias. The result of applying a spatial scan statistic based on echelon analysis was the detection of a significant space-time cluster that decreased with the passing of time. Moreover, the detected cluster was located in the direction of northwest from the NPS.

CC0401: **Cluster detection of disease mapping data based on latent Gaussian Markov random field models**
Presenter: **Wataru Sakamoto**, Okayama University, Japan

Detecting clusters of higher prevalence in spatial data is of primary interest. Most of the existing methods use spatial scan statistics based on the likelihood-ratio test. The echelon scan based on the echelon analysis is useful in detecting significant clusters of non-circular shape effectively. Bayesian analysis methods for spatial data have been also studied. The latent Gaussian models, in which a Gaussian Markov random field prior is assumed on the spatial effect, provide very flexible tools. A method of detecting clusters was proposed using Poisson models with the latent Gaussian Markov random field. The clusters are scanned on the echelons constructed from the posterior means of the spatial effect, and the clusters giving maximum marginal likelihood were detected. It can be easily extended to adjustment for covariates and the random effect. An example of applying to disease mapping data illustrated that the proposed method constructed more aggregated echelons and clusters than the echelon scan based on empirical Bayes estimates of relative risk, and that detected clusters provided smallest deviance information criterion values.

CC0221: **ProbitSpatial R package: Fast and accurate spatial probit estimations**
Presenter: **Davide Martinetti**, INRA, France
Co-authors: Ghislain Geniaux

This package meets the emerging needs of powerful and reliable models for the analysis of spatial discrete choice data. Since the explosion of available and voluminous geospatial and location data, older estimation techniques cannot withstand the course of dimensionality and are restricted to samples with less than a few thousand observations. The functions contained in ProbitSpatial allow fast and accurate estimations of Spatial Autoregressive and Spatial Error Models under Probit specification. They are based on the full maximization of likelihood of an approximate multivariate normal distribution function, a task that was considered as prodigious just few years ago. Extensive simulation and empirical studies proved that these functions can readily handle sample sizes with as many as several millions of observations, provided the spatial weight matrix is in convenient sparse form, as is typically the case for large data sets, where each observation neighbours only a few other observations. SpatialProbit relies amongst others on Rcpp, RcppEigen and Matrix packages to produce fast computations for large sparse matrixes. Possible applications of spatial binary choice models include spread of diseases and pathogens, plants distribution, technology and innovation adoption, deforestation, land use change, amongst many others.

CC0465: **A framework for spatio-temporal regression analysis of extremes in a presence of missing covariates**
Presenter: **Olga Kaiser**, Universita della Svizzera italiana, Switzerland
Co-authors: Illia Horenko

Statistical regression analysis of extreme events aims to describe their behavior as a relationship between some covariates and the observed extremes. Often the complete set of all the significant covariates can not be provided. In such a case nonstationary and nonparametric approaches are required to obtain unbiased results. Based on the Generalized Extreme Value distribution (GEV) and the nonparametric Finite Element time-series analysis Methodology (FEM) with Bounded Variation on the model parameters (BV) we build a semiparametric, nonstationary, and non-homogenous computational framework for regression analysis of spatio-temporal extremes. The FEM-BV-GEV framework addresses the issue of nonstationarity with describing the underlying dynamics by $K \geq 1$ local stationary models and a nonstationary, nonparametric and nonhomogenous switching process. The switching process can provide insights into the pattern of the systematically missing covariates. We show cases when the analysis of the switching process reveals explicitly the missing covariate. Further, the framework provides a spatio-temporal clustering of extreme events and so a pragmatic, nonparametric and nonstationary description of the underlying spatial dependence structure. The performance of the framework is demonstrated on monthly maximum temperature data over Europe.

APPLIED STATISTICS AND ECONOMETRICS

CC064 **Room Sala 7** **Chair: Peter Winker**

CC0434: **Monetary policy on Twitter and its effect on asset prices: Evidence from computational text analysis**
Presenter: **Peter Tillmann**, Justus-Liebig University Giessen, Germany
Co-authors: Jochen Luedering

The aim is to dissect the public debate about the future course of monetary policy and trace the effects of selected topics of this discourse on U.S. asset prices. We focus on the "taper tantrum" episode in 2013, a period with large revisions in expectations about Fed policy. Based on a novel data set of 90,000 Twitter messages ("tweets") covering the entire debate of Fed tapering on Twitter we use Latent Dirichlet Allocation, a computational text analysis tool to quantify the content of the discussion. Several estimated topic frequencies are then included in a VAR model to estimate the effects of topic shocks on asset prices. We find that the discussion about Fed policy on social media contains price-relevant information. Shocks to shares of "tantrum"-, "QE"- and "data"-related topics are shown to lead to significant asset price changes. We also show that the effects are mostly due to changes in the term premium of yields consistent with the portfolio balance channel of unconventional monetary policy.

CC0469: **Measuring dependence between dimensions of welfare using multivariate Spearman's rho and other copula-based coefficients**
Presenter: **Ana Perez Espartero**, University of Valladolid, Spain
Co-authors: Mercedes Prieto

Welfare is multidimensional as it involves not only income, but also education, health or labor. The composite indicators of welfare are usually based on aggregating somehow the information across dimensions and individuals. However, this approach ignores the relationship between the dimensions being aggregated. To face this goal, we analyse the multivariate dependence between the dimensions included in the Human Development Index (HDI), namely income, health and schooling, through three copula-based measures of multivariate association: Spearman's

footrule, Gini's gamma and Spearman's rho. We discuss their properties and prove new results on Spearman's footrule. The copula approach focuses on the positions of the individuals across dimensions, rather than the values that the variables attain for such individuals. Thus, it allows for more general types of dependence than the linear correlation. We base our study on data from 1980 till 2014 for the countries included in the 2015 Human Development Report. We find out that though the overall HDI has increased over this period, the dependence between the dimensions included in the HDI (income, health and schooling) remains high and nearly unchanged so that the richest countries tend to be also the best ranked in both health and education.

CC0414: **Joint modeling of inflation and real interest rate dynamics with application to equity-linked investment**

Presenter: **Arto Luoma**, University of Tampere, Finland

Co-authors: Lasse Koskinen, Tommi Salminen

A realistic model is introduced for the joint dynamics of real interest rate and inflation so that it could be used for various prediction purposes, for example to analyze the role of inflation in the pricing and hedging of financial derivatives. In a combined auto-regressive process, normal or more stable inflation periods are explained by a standard AR-process, while unanticipated peaks are captured by an additional process following Student's t-distribution. Next, the effect of inflation on the pricing of an equity index linked insurance product is studied. The models are estimated using Bayesian methods with US data.

CC0172: **Nonlinear dynamics and wavelet based analysis of crude oil prices**

Presenter: **Emmanuel Senyo Fianu**, Recent Affiliation–Leuphana University of Lueneburg, Germany

A signal modality analysis is investigated for the characterization and the detection of nonlinearity in crude oil markets. Given the nonlinear and time-varying characteristics of international crude oil prices, this analysis employs the recently proposed delay vector variance (DVV) method, which examines local predictability of a signal in the phase space to detect the determinism and nonlinearity in the energy time series. In addition, wavelet transforms, which have recently emerged as a mathematical tool for multiresolution decomposition of signals, have been employed. In particular, a complex Morlet wavelet is used to detect and characterize the various phases of oil prices through the trajectory of its evolution. It has the potential applications in signalling processing that require variable time-frequency localizations. A detail overview of the feasibility of this methodology is highlighted. Our results aims at identifying the significant phases of the series and its relation to real-world phenomena in recent years as well future occurrences.

Thursday 25.08.2016 16:45 - 18:35 Parallel Session M – COMPSTAT

RECENT ADVANCES ON FUNCTIONAL DATA ANALYSIS AND APPLICATIONS

CO012 **Room Sala 1** **Chair: Ana M Aguilera**

CO0187: **Multivariate functional principal component analysis for data observed on different (dimensional) domains**
Presenter: **Clara Happ**, LMU Munich, Germany
Co-authors: Sonja Greven

Existing approaches for multivariate functional principal component analysis are restricted to data on a single interval $\mathcal{T} \subset \mathbb{R}$. The presented approach focuses on multivariate functional data $X = (X^{(1)}, \ldots, X^{(p)})$ observed on different domains $\mathcal{T}_1, \ldots, \mathcal{T}_p$ that may differ in dimension, e.g. functions and images. The theoretical basis for multivariate functional principal component analysis is given in terms of a Karhunen-Loeve theorem. For the practically relevant case of a finite, possibly truncated, Karhunen-Loeve representation, a direct theoretical relationship between univariate and multivariate functional principal component analysis is established. This offers a simple estimation strategy to calculate multivariate functional principal components and scores based on their univariate counterparts. The approach can be extended to univariate components $X^{(j)}$ that have a finite expansion in a general, not necessarily orthonormal basis and is applicable for sparse data or data with measurement error. A flexible software implementation for representing multivariate functional data and estimating the multivariate functional PCA is made available in two R-packages. The approach is applied to a neuroimaging study to explore how longitudinal trajectories of a neuropsychological test score covary with FDG-PET brain scans at baseline.

CO0206: **An ANOVA test for functional data with graphical interpretation**
Presenter: **Tomas Mrkvicka**, University of South Bohemia, Czech Republic
Co-authors: Mari Myllymaki, Ute Hahn

A new functional ANOVA test, with a graphical interpretation of the result, will be presented. The test is an extension of a global envelope test recently introduced. The graphical interpretation is realized by a global envelope which is drawn jointly for each sample of functions. If an average function, computed over a sample, is out of the given envelope, the null hypothesis is rejected with the predetermined significance level α. The advantages of the proposed procedure are that it identifies the domains of the functions which are responsible for the potential rejection and that it immediately offers a post-hoc test by identifying the samples which are responsible for the potential rejection. All that is done at an exact significance level α. Our simulations show that the power of the test also tends to be higher than the power of recently available procedures.

CO0209: **Bootstrap confidence intervals in semi-functional partial linear regression under dependence**
Presenter: **Paula Rana Miguez**, Universidade da Coruna, Spain
Co-authors: German Aneiros-Perez, Juan Vilar Fernandez, Philippe Vieu

Two bootstrap procedures, naive and wild bootstrap, are proposed to construct pointwise confidence intervals for the semi-functional partial lineal regression model, when the response is scalar and considering scalar covariates (for the linear component) and functional covariates (for the nonparametric component). By means of these two bootstrap procedures, we can approximate the asymptotic distribution for both parts in the model: the linear and the nonparametric components. The validity of the two bootstrap procedures has been proved theoretically in the setting of dependent data, assuming alpha-mixing conditions on the sample, and also for independent data as a particular case. Naive bootstrap allows dealing with homoscedastic data, meanwhile wild bootstrap is devoted to work with heteroscedastic data. Pointwise confidence intervals for each component of the model have been built. A simulation study shows the performance of the procedure, which has been also applied to a real dataset. Specifically, an application to electricity price from the Spanish Electricity Market illustrates its usefulness in practice.

CO0291: **Multi-classification of human body motions from acceleration curves**
Presenter: **M Carmen Aguilera-Morillo**, Universidad Carlos III de Madrid, Spain
Co-authors: Ana M Aguilera

The aim is to classify a set of functional data according to a multinomial response variable. Exactly, a set of accelerations curves measured during the realization of three different body motions (walking, walking upstairs and walking downstairs) have been considered. In order to solve this multi-classification problem, methodology based on functional linear discriminant analysis (FLDA) of the multinomial response variable on a set of functional PLS components of the acceleration curves is proposed. With the aim of improving both, the classification of new samples curves and the estimation of the discriminant functions, two penalized versions of functional PLS regression are combined with FLDA. The results obtained from this dataset highlight the need to use a penalization term, in which case a correct classification rate greater than the 80% has been achieved on the test sample.

CO0348: **Reconstructing gradients from sparse functional data**
Presenter: **Ian McKeague**, Columbia University, United States

Consider the problem of estimating growth curves $X_j(t), j = 1, \ldots, N$ over a time interval $[0, T]$, from data on their integrals over gaps between n observation times. We introduce a new Bayesian approach to this problem taking into account: (1) each trajectory X_j is known to be uniformly bounded, (2) there is measurement error in the data, and (3) pairs of trajectories are unlikely to cross very often. To address (1), we make use of a hyperbolic-tangent transform applied to tied-down Brownian motion, as well as a multivariate normal at observation times, to form a natural bounded process for use as a prior. To address the measurement error issue (2), the log-transformed data are modeled with additive Gaussian noise. To address (3), we use N non-intersecting tied-down Brownian bridges to provide an ensemble prior between the observation times.

SURVEY SAMPLING

CO093 **Room Sala 2** **Chair: Alina Matei**

CO0287: **Coordination of spatial samples**
Presenter: **Alina Matei**, University of Neuchatel, Switzerland
Co-authors: Anton Grafstrom

Sample coordination seeks to create a probabilistic dependence between two or more samples. The goal is to maximize the number of common units (positive coordination) to improve the estimation precision or to minimize this number (negative coordination) to reduce the nonresponse rate of units involved in different surveys. The sample coordination is applied in spatial sampling and it is achieved using permanent random numbers. The goal is to coordinate spatial samples, while preserving their good spatial properties. The overall approach is motivated by examples from official statistics and forestry.

CO0181: **Estimation of the finite population distribution function using a global penalized calibration method**
Presenter: **Maria Dolores Jimenez-Gamero**, University of Sevilla, Spain
Co-authors: Jose Antonio Mayor-Gallego, Juan Luis Moreno-Rebollo

An estimator of the finite population distribution function $F_y(t)$ when auxiliary variables x, related to the study variable y by a superpopulation model, are available, is proposed. The new estimator integrates ideas from model calibration, and penalized calibration. Alternatively to other model based calibration estimates of $F_y(t)$ in the literature, which require the distribution function estimation of the fitted values to be equal to the

distribution function of the fitted values on some fixed points, a penalty term that measures the distance between the distribution function of the fitted values and its estimate, is included in the objective function. Thus, in a sense, it is imposed on both distribution functions be close at all points. Conditions are given so that the proposed estimate to be a proper distribution function. Results on the asymptotic unbiasedness and the asymptotic variance of the proposed estimator are obtained. In a simulation study the proposed estimator has better performance than others in the literature.

CO0271: **A systematic approach for choosing a sampling design**
Presenter: **Yves Tille**, University of Neuchatel, Switzerland

The selection of a sampling design is a complex procedure that depends on the knowledge that we have of a population of interest. We propose a methodology that consists of modeling the populations and next of choosing the sampling design in function of the selected model. The choice of the sampling design is an arbitration between three main principles: the principle of randomization, the principle of restriction and the principle of overrepresentation.

CO0311: **Small area estimation: A nonparametric maximum likelihood approach**
Presenter: **Maria Francesca Marino**, University of Perugia, Italy
Co-authors: Maria Giovanna Ranalli, Nicola Salvati, Marco Alfo

When dealing with small area estimation, generalized linear mixed models represent a typical frequentist tool for deriving best prediction of counts or proportions. Area-specific Gaussian random parameters are typically considered to account for sources of unobserved heterogeneity that are not captured by the covariates in the model. However, for non-Gaussian responses, computing the EBP and the corresponding MSE requires the solution of (possibly) multiple integrals that do not admit closed form. Monte Carlo methods and parametric bootstrap are frequent choices even if the computational burden represents a non trivial task. We propose to estimate model parameters via a nonparametric maximum likelihood approach (NPML). We derive the EBP and the analytic approximation to its MSE. NPML allows us to avoid unverifiable assumptions on the random parameter distribution: as long as the likelihood is bounded, its maximization leads to a finite mixture distribution with at most as many support points as the number of distinct area profiles. Also, since mixture parameters are directly estimated from the data, extreme and/or asymmetric departures from the homogeneous model can be accommodated. Last, given the discrete nature of the mixing distribution, we can avoid integrals approximation and Monte Carlo integration thus considerably reducing the computational effort.

CO0285: **Estimation of total electricity consumption curves of small areas by sampling in a finite population.**
Presenter: **Anne de Moliner**, Universite de Bourgogne EDF, France
Co-authors: Herve Cardot, Camelia Goga

Many studies carried out in the French electricity company EDF are based on the analysis of the total electricity consumption curves of groups of customers. These aggregated electricity consumption curves are estimated by using samples of thousands of curves measured at a small time step and collected according to a sampling design. Small area estimation is very usual in survey sampling. It is often addressed by using implicit or explicit domain models between the interest variable and the auxiliary variables. The goal is to estimate totals of electricity consumption curves over domains or areas. Three approaches are compared: the first one consists in modeling the functional principal scores with linear mixed models. The second method consists in using functional linear regression models and the third method, which is non-parametric, is based on regression trees for functional data. These methods are evaluated on a dataset of French consumption curves.

TUTORIAL 3

CO104 **Room Sala Camara** **Chair: Alessandra Amendola**

CO0581: **Band pass filtering and wavelets analysis**
Presenter: **Stephen Pollock**, University of Leicester, United Kingdom

Wavelets analysis provides a means of analysing non-stationary time series of which the underlying statistical structures are continually evolving. It is an analysis both in the time domain and in the frequency domain. The effects of digital filtering in the time domain and the frequency domain will be described. It will proceed to provide the generalisation of the Shannon sampling theorem that is appropriate to bandpass filtering. This theorem establishes a relationship between continuous signals and their corresponding sampled sequences that is essential to a wavelets analysis. Once this background has been provided, the theories of Dyadic and non-Dyadic wavelets analysis can be described in detail.

LATENT VARIABLE MODELS

CG062 **Room Sala 3** **Chair: Sara Taskinen**

CC0194: **Evaluation of the robustness of stepwise latent class estimators and a new two-stage estimator**
Presenter: **Zsuzsa Bakk**, Leiden university, Netherlands

The aim focuses on classification error corrected stepwise estimation approaches of Latent Class (LC) models with external variables. Currently two approaches are available, ML and BCH, that follow a similar procedure: in the first step the underlying latent construct is estimated based on a set of observed indicator variables, next, in step two, cases are assigned to the LCs, and, finally, in the third step, the class assignments are used in further analyses, while correcting for classification error. We discuss the robustness of the stepwise estimation procedures to different violations of underlying model assumptions and highlight that when the presence of direct effects between the external variable and the indicators of LC model are ignored, both approaches lead to biased estimates in the last step. We propose an alternative two-stage estimator to address this problem. Using this two-stage estimator in step one, a LC model is estimated with the indicators only, and in the second step, the external variable of interest is added to the model, freely estimating the association between the external variable and LC membership, while keeping the parameters of the measurement model fixed to the values estimated in step one. Using this approach, local fit measures can be used to test which fixed effects need to be freed to account for the presence of direct effects.

CC0427: **Using forward search algorithm for person fit analysis in general cognitive diagnosis models**
Presenter: **Kevin Carl Santos**, University of the Philippines, Philippines
Co-authors: Jimmy de la Torre, Erniel Barrios

Cognitive diagnosis models (CDMs) are psychometric models used to identify the strengths and weaknesses of examinees based on multidimensional attribute patterns. However, these latent attribute profiles may be inaccurately estimated because of an atypical test performance reflected in the response patterns. Person fit assessment is then performed to obtain information regarding the aberrant behavior of the examinees. The aim is to extend the forward search algorithm to general CDMs, particularly the G-DINA (generalized deterministic inputs, noisy and gate) model, to identify aberrant response patterns. Methods to select the initial set, and to progress and monitor the search are explored. Forward plots of goodness-of-fit statistics and Cooks distance are examined to observe drastic changes. A simulation study is conducted to determine the performance of the proposed method on different scenarios.

CC0494: **Approximate likelihood inference via dimension reduction in latent variable models for categorical data**
Presenter: **Silvia Cagnone**, University of Bologna, Italy
Co-authors: Silvia Bianconcini, Dimitris Rizopoulos

Latent variable models represent a useful tool in different fields of research in which the constructs of interest are not directly observable, so

that one or more latent variables are required to reduce the complexity of the data. In these cases, problems related to the integration of the likelihood function of the model can arise since analytical solutions do not exist. Usually, numerical quadrature-based methods like Gauss-Hermite or adaptive Gauss-Hermite are used to overcome this problem. They work quite well in several situations, but become unfeasible in presence of many latent variables and/or random effects. We propose a new approach, referred to as Dimension Reduction Method (DRM), that consists of a dimension reduction of the multidimensional integral that makes the computation feasible in situations in which the quadrature based methods are not applicable. We discuss the advantages of DRM compared with other existing approximation procedures in terms of both computational feasibility as well as asymptotic properties of the resulting estimators. Applications to real data are also illustrated.

CC0440: **Penalized latent variable models**
Presenter: **Brice Ozenne**, Copenhagen University, Denmark
Co-authors: Klaus Kahler Holst, Esben Budtz-Jorgensen

Latent variables models (LVM) are statistical models able to relate measurements in a very flexible, and possibly complex, way. However they are not suited to study high dimensional data that arise, for instance, in genetics or in medical imaging. Moreover variable selection within LVM currently relies on stepwise testing procedures that suffer from instability. We propose to extend the Gaussian LVM to allow penalization on the mean or covariance parameters. An elastic net penalty is used for the mean parameter; this penalization includes lasso and ridge penalization as specific cases. A group lasso is used for penalizing the covariance structure. Estimation of the model relies on a proximal gradient algorithm to handle the non-derivability of the lasso penalty. The ability of the penalized LVM to identify the relevant parameters will be investigated in simulation studies including high dimensional settings. We will then assess its relevance for relating measurements of the serotonin in the human brain to the depression status of patients. The penalized LVM is implemented as an add-on of the R package lava and is available upon request.

CC0535: **Generalized linear latent variable models for analyzing multivariate abundance data**
Presenter: **Sara Taskinen**, University of Jyvaskyla, Finland
Co-authors: Jenni Niku, Francis Hui, David Warton

In ecological studies, abundances of many, interacting species are often collected in several sites. Such data are often very sparse, high-dimensional and include highly correlated responses. The main aim of the statistical analysis is then to understand the interrelationships among such multiple, correlated responses. We consider model-based approaches for analyzing multivariate abundance data. We will show how generalized linear latent variable models (GLLVMs) can easily capture the correlation inherent in responses and provide a powerful tool for estimation and inference. Fast and efficient maximum likelihood based algorithms for fitting the models will be discussed. It is shown that especially variational approximation method performs better than several classical estimation methods for GLLVMs. The methods will be applied to ecological datasets focusing on model-based approaches to unconstrained ordination.

MULTIVARIATE DATA ANALYSIS

CC069 **Room Sala 5** **Chair: Christian Hennig**

CC0530: **Affine invariant algorithms for nonnegative matrix and 3D tensor factorization**
Presenter: **Ruoni Zhang**, Hokkaido University, Japan
Co-authors: Hideyuki Imai

A novel model of NMF is proposed to incorporate an affine transformation and their extensions to 3D tensor factorization. It can deal with the mixed signs and the factorization results hold the property of affine invariant. Multiplicative estimation algorithms are also provided for the resulting this model.

CC0570: **The Matsumoto-Yor property on trees for matrix variates of different dimensions**
Presenter: **Konstancja Bobecka**, Warsaw University of Technology, Poland

The focus is on an extension of the multivariate Matsumoto-Yor (MY)independence property with respect to a tree with p vertices to the case where random variables corresponding to the vertices of the tree are replaced by random matrices. The converse of the $p-$variate MY property, which characterizes the product of one gamma and $p-1$ generalized inverse Gaussian distributions, is extended to characterize the product of the Wishart and $p-1$ matrix generalized inverse Gaussian distributions.

CC0488: **Test for covariance structure for high-dimensional data under non-normality**
Presenter: **Takahiro Nishiyama**, Senshu University, Japan
Co-authors: Yuki Yamada, Masashi Hyodo

A test is proposed for making an inference about the block-diagonal covariance structure of a covariance matrix in non-normal high-dimensional data. Since the classical hypothesis testing methods based on the likelihood ratio degenerate when the dimensionality exceeds the sample size, we instead turn to a distance function between the null and alternative hypothesis. We prove that the limiting null distribution of the proposed test is normal under mild conditions when its dimension is substantially larger than its sample size. We further study the local power of the proposed test. Finally, we study the finite sample performance of the proposed test via Monte Carlo simulations. We demonstrate the relevance and benefits of the proposed approach for a number of alternative covariance structures.

CC0220: **Analysis of rotational deformations from directional data using a parametric and non-parametric approach**
Presenter: **Joern Schulz**, University of Stavanger, Norway
Co-authors: Byung-Won Kim, Stephan Huckemann, Steve Marron, Stephen Pizer, Sungkyu Jung

Rotational deformations such as bending or twisting have been observed as the major variation in various medical applications. To provide a better surgery or treatment planning, it is crucial to model such deformations of 3D objects that can be described by the movements of directional vectors on the unit sphere. Such multivariate directional vectors are available in a number of different object representations and the rotation of each vector follows a small circle on the unit sphere. Thus, the proposed parametric and non-parametric estimation procedures are based on small circles on the unit sphere. The parametric approach is a likelihood-based estimation procedure using two novels small circle distributions called the Bingham-Mardia Fisher distribution and Bingham-Mardia-von Mises distribution. The proposed estimation procedures can model dependence structure between directions and facilitate hypotheses testing. In the non-parametric approach, estimates of the rotation axis and angles are obtained by fitting small circles applying sample Frechet means and least-square estimators. The performance of the proposed estimators are demonstrated i) in a simulation study, ii) on deformed ellipsoids and iii) on knee motions during gait.

CC0561: **Components selection for multivariate outlier detection with ICS**
Presenter: **Aurore Archimbaud**, Toulouse School of Economics, France
Co-authors: Anne Ruiz-Gazen, Klaus Nordhausen

The detection of a small proportion of multivariate outliers such as identifying production errors in industrial processes is an important topic. In this context, the Invariant Coordinate Selection (ICS) method is an efficient identification procedure. The ingenious idea of the method, compared to other multivariate methods such as Principal Component Analysis (PCA) or robust PCA, is to simultaneously diagonalize two scatter matrices. In case of a small percentage of outliers, the ICS coordinates are ordered decreasingly according to a generalized concept of kurtosis depending on the considered pair of scatters. Taking into account the coordinates associated with large kurtosis values, the observations far away from the center of the data are declared as outliers. One challenging step in the procedure is to select the components that display outliers. Two approaches are

introduced and compared. The first one is comparable to a test procedure where the critical value is calculated using some simulations. The other approach incorporates some univariate normality tests. The first approach is time consuming and depends on the sample size, the dimension and the pair of scatters involved in ICS. Some approximations are thus constructed in order to avoid to carry out new simulations for each case.

MIXTURE MODELS		
CG015	**Room Sala 4**	**Chair: Geoffrey McLachlan**

CC0388: **On multivariate extensions of the Mixed Tempered Stable distribution**
Presenter: **Asmerilda Hitaj**, Milano Bicocca, Italy
Co-authors: Friedrich Hubalek, Lorenzo Mercuri, Edit Rroji

A generalization of Normal Variance Mean Mixtures, named multivariate Mixed Tempered Stable distribution, is studied. Properties of this distribution and its capacity in capturing fat tails are discussed based on simulation analysis. We point out that this distribution is suitable in reproducing stylized facts and different dependence structures between asset returns.

CC0323: **Asymptotic normality of the maximum likelihood estimator for the latent block model**
Presenter: **Christine Keribin**, INRIA - Universite Paris-Sud, France
Co-authors: Vincent Brault, Mahendra Mariadassou

Latent Block Model (LBM) is a probabilistic method based on a mixture model to cluster simultaneously the d columns and n rows of a data matrix. Maximum likelihood parameter estimation in LBM is a difficult and multifaceted problem, as neither the likelihood, nor the expectation of the conditional likelihood are numerically tractable. Then, the standard EM must be adapted and various estimation strategies have been proposed and are now well understood empirically. But as far as now, theoretical guarantees about their asymptotic behaviour is rather sparse. We show here that under some mild conditions on the parameter space, and in an asymptotic regime where $\log(d)/n$ and $\log(n)/d$ go to 0 when n and d go to $+\infty$, (1) the maximum-likelihood estimate of the complete model (with known labels) is consistent and (2) the log-likelihood ratios are equivalent under the complete and observed (with unknown labels) models. This equivalence allows us to transfer the asymptotic consistency (1) to the maximum likelihood estimate under the observed model.

CC0183: **A classical invariance approach to the normal mixture problem**
Presenter: **Monia Ranalli**, The Pennsylvania State University, United States
Co-authors: Bruce Lindsay, David Hunter

Although normal mixture models have received great attention and are commonly used in different fields, they stand out for failing to have a finite maximum to the likelihood. In the univariate case there are n solutions, corresponding to the n distinct data points, along a parameter boundary, each with an infinite spike of the likelihood, none making particular sense as a chosen solution. In the multivariate case, there is an even more complex likelihood surface. We show that there is a marginal likelihood that is bounded and quite close to the full likelihood in information as long as one is interested in the central part of the parameter space, away from its problematic boundaries. Our main goal is to show that the marginal likelihood solves the unboundedness problem and in a manner competitive with other methods that were specifically designed for the normal mixture. To this aim, two different algorithms have been developed. Their effectiveness is investigated through a simulation study. Finally, an application to real data is illustrated.

CC0464: **Accounting for the sparsity rate in prior selection in inverse problems via generalized Student$-t$ distribution**
Presenter: **Li Wang**, CNRS - Universite Paris-Sud - CentraleSupelec, France
Co-authors: Mircea Dumitru, Ali Mohammad-Djafari

Bayesian methods have become very common for inverse problems arising in signal and image processing. The main advantages are the possibility to propose unsupervised methods where the likelihood and prior model parameters can be estimated jointly with the main unknowns and to select prior distributions that are in accordance with the prior knowledge. In the context of sparsity enforcing priors, the selection of the prior distribution can be done also in accordance with the sparsity rate of the unknown of the model. This can be achieved based on the generalization of the Student$-t$ distribution. First, the generalization of the Student$-t$ distribution is proposed, based on its Infinite Gaussian Scaled Mixture (IGSM) model. The generalization is obtained as the marginal posterior distribution of the mean of a Gaussian distribution with unknown variance on which an a priori Inverse Gamma distribution is assigned. Then, some of its properties are discussed, namely the computation of the variance and its interpretation in the context of accounting for the sparsity rate and the links with the corresponding Inverse-Gamma distribution in the context of sparsity enforcing mechanism in the Bayesian approach. Simulations results and comparisons are presented for applications in Computed Tomography and chronobiology.

TIME SERIES		
CC072	**Room Sala 7**	**Chair: Ana Perez Espartero**

CC0207: **Change point detection by filtered derivative with p-Value: Choice of the extra-parameters**
Presenter: **Doha Hadouni**, Blaise Pascal, France
Co-authors: Pierre Bertrand

The study deals with off-line change point detection using the Filtered Derivative with p-Value method. The FDpV method is a two-step procedure for change point analysis. The first step is based on the Filtered Derivative function (FD) to select a set of potential change points, using its extra-parameters - namely the threshold for detection and the sliding window size. In the second step, we compute the p-value for each change point in order to retain only the true positives and discard the false positives. We give a way to estimate the optimal extra-parameters of the function FD, in order to have the fewest possible false positives and non-detected change points. Furthermore, we give the threshold of the p-value such that we detect only the real change points. Indeed, the estimated potential change points may differ slightly from the theoretically correct ones. After setting the extra-parameter in the two steps, we need to know whether the absence of detection or the false alarm has more impact on the Mean Integrated Square Error, which prompts us to calculate the MISE in both cases. Finally, a simulation study with a Monte Carlo method and the applications on the real data of heart-rate beat show the positive and negative ways the parametrisation can affect the results.

CC0441: **Modeling bivariate count series through dynamic factor models**
Presenter: **Magda Monteiro**, University of Aveiro, Portugal
Co-authors: Isabel Pereira, Manuel Scotto

Current research on count series modeling has its focus centered on multivariate models. These models either belong to the class of observation driven model or to the class of parameter driven model. Belonging to the former class is one of the first multivariate count model using a multivariate Poisson state space model. The work was later generalized by proposing a dynamic factor model for multivariate count data which allow for temporal and contemporaneous interaction between series. The aim is to present a dynamic factor model for bivariate count series whose mean vector depends on an autoregressive component beyond a common latent factor. An application of this model is made to fire activity, namely, to the monthly number of forest fires in the neighboring districts of Aveiro and Coimbra, Portugal. We use a Bayesian approach, through MCMC methods, to estimate the model parameters as well as to estimate the common factor to both series.

CC0533: **Sliced inverse regression for time series**

Presenter: **Markus Matilainen**, University of Turku, Finland
Co-authors: Christophe Croux, Klaus Nordhausen, Hannu Oja

When analysing data with a response variable y and explanatory variables $\mathbf{x}$, modelling may become infeasible when the number of variables gets higher. It can also cause computational problems and visualization of data becomes harder. To avoid these kind of problems Sliced Inverse Regression (SIR) can be used. It is used to find the subspace of $\mathbf{x}$ which contains all the essential information needed to model y. However, SIR was developed for iid data. An extension for SIR where both $\mathbf{x}_t$ and y_t are time series is suggested. The new method uses several supervised lagged autocovariance matrices and can then also indicate which lags of $\mathbf{x}_t$ are relevant to the modelling process of y_t. Different ways to choose the lags and the number of dimensions to keep are suggested. The method and different selection ways are demonstrated using simulated and real data.

CC0161: **Calculating joint confidence bands for impulse response functions using highest density regions**

Presenter: **Peter Winker**, University of Giessen, Germany
Co-authors: Helmut Luetkepohl, Anna Staszewska-Bystrova

A new non-parametric method is proposed to construct joint confidence bands for impulse response functions of vector autoregressive models. The estimation uncertainty is captured by means of bootstrapping and the highest density region (HDR) approach is used to construct the bands. A Monte Carlo comparison of the HDR bands with existing alternatives shows that the former are competitive with the bootstrap-based Bonferroni and Wald confidence regions. The relative tightness of the HDR bands matched with their good coverage properties makes them attractive for applications. An application to corporate bond spreads for Germany highlights the potential for empirical work.

CC0515: **Randomized singular spectrum analysis for long time series**

Presenter: **Paulo Canas Rodrigues**, Federal University of Bahia, Brazil
Co-authors: Petala Tuy, Rahim Mahmoudvand

Singular spectrum analysis (SSA) is a relatively new and powerful nonparametric method for analyzing time series that is an alternative to the classic methods. This methodology has proved to provide an efficient analysis of time series in various disciplines as the assumptions of stationarity and Gaussian residuals can be relaxed. The Era of Big Data has brought very long and complex time series. Although SSA have provided advantages over traditional methods, the computational time needed for the analysis of long time series might make it unappropriated. We propose the randomized SSA which intends to be an alternative to SSA for long time series without losing the quality of the analysis. The SSA and the randomized SSA are compared in terms of quality of the analysis and computational time, using Monte Carlo simulations and real data.

Friday 26.08.2016 09:00 - 10:30 Parallel Session N – COMPSTAT

ALGORITHMS FOR CATEGORICAL DATA

CI079 **Room Sala Camara** **Chair: Tamas Rudas**

CI0270: **On variants of the iterative scaling algorithm**
Presenter: **Tamas Rudas**, Eotvos Lorand University, Hungary
Co-authors: Anna Klimova

The Iterative Scaling Algorithm has been present in statistics for a long time and is routinely used in many applications, including small area estimation in official statistics, post-stratification in survey analysis and maximum likelihood estimation of log-linear models. The main focus of interest is the latter area of application. First, a unified treatment of the original IPS and two of its modifications, the Generalized Iterative Scaling, and of the Improved Iterative Scaling, are given. Then, it is shown that these algorithms cannot deal with the problem of maximum likelihood estimation in a coordinate-free generalization of the log-linear model, called relational models. In these models, the sample space does not have to be a Cartesian product, the multiplicative effects are not necessarily associated with cylinder sets and an overall effect may not be present. Maximum likelihood estimates have many surprising properties. A new variant of IS is described, which may be applied to relational models, and it is shown that a generalization of it also works in the presence of zero observed frequencies.

CI0277: **Mixed parametrization, IPF and fixed point algorithms in marginal models**
Presenter: **Antonio Forcina**, Perugia, Italy

Distributions belonging to the exponential family may be parameterized by combining an arbitrary selection of mean parameters with the complementary set of canonical parameters. Within the multinomial distribution, mean parameters are marginal probabilities while canonical parameters are log-linear contrasts. A Newton algorithm may be used to reconstruct the joint distribution with a given mixed parametrization. Because any set of log-linear parameters are variation independent and the mean parameters are also variation independent from the canonical parameters, the corresponding joint distribution will always exist as long as the mean parameters are internally consistent. The reconstruction algorithm may be seen as an alternative to IPF for fitting log-linear models by setting the mean parameters to the observed marginals and the complementary set of log-linear parameters to 0 (or to arbitrary values). The same algorithm may be used to show that a collection of log-linear parameters defined within different marginal distributions are smooth when marginals may be ordered so that each one is parameterized by a set of log-linear parameters combined with the mean parameters available from preceding marginals. Arguments based on fixed point algorithms combined with the mixed parametrization have been used to prove smoothness of more complex marginal parameterizations.

CI0310: **A unified approach to marginal and conditional independencies of binary variables based on Moebius inversion**
Presenter: **Luca La Rocca**, University of Modena and Reggio Emilia, Italy
Co-authors: Alberto Roverato

A novel parameterization is presented for the joint distribution of a binary vector, based on partitioning the vector in two sets of variables. This parameterization, which we name hybrid parameterization, provides us with a unified expression for all conditional and marginal independencies implied by a class of bipartite regression graphs. Both undirected and bidirected graphical models belong to this class, as extreme cases, and the hybrid parameterization specializes to the traditional log-linear parameterization and the more recent log-mean linear parameterization, respectively, in these cases. We illustrate the role the hybrid parameterization can play in the study of relationships among binary variables.

RECENT CONTRIBUTIONS TO ROBUST MIXTURE MODELLING

CO095 **Room Sala 1** **Chair: Francesca Greselin**

CO0254: **Robust mixture modeling by mean shift parameters**
Presenter: **Weixin Yao**, UC Riverside, United States
Co-authors: Chun Yu, Kun Chen

Finite mixture regression models have been widely used for modelling mixed regression relationships arising from a clustered and thus heterogenous population. The classical normal mixture model, despite of its simplicity and wide applicability, may fail in the presence of severe outliers. We propose a new robust mixture regression approach based on a sparse, case-specific, and scale-dependent mean-shift mixture model parameterization, for simultaneously conducting outlier detection and robust parameter estimation. A penalized likelihood approach is adopted to induce sparsity among the mean-shift parameters so that the outliers are distinguished from the good observations, and a thresholding-embedded Expectation-Maximization (EM) algorithm is developed to enable stable and efficient computation. The proposed penalized estimation approach is shown to have strong connections with other robust methods including the trimmed likelihood method and the M-estimation approaches. Comparing with several existing methods, the proposed methods show outstanding performance in our numerical studies.

CO0358: **Robust clustering of multivariate skew data**
Presenter: **Francesca Greselin**, University of Milano Bicocca, Italy
Co-authors: Luis Angel Garcia-Escudero, Agustin Mayo-Iscar, Geoffrey McLachlan

With the increasing availability of multivariate datasets, attention is being directed to providing more robust methods than classical approaches for model based clustering like mixtures of Gaussian distributions. Moreover, in performing ML estimation we know that a few outliers in the data can affect the estimation, hence providing unreliable inference. Challenged by such issues, more flexible and solid tools for modeling heterogeneous skew data are needed. We introduce a robust approach for estimating mixtures of canonical fundamental skew normal, based on trimming outlying observations and performing constrained estimation. We also provide a feasible EM to implement model estimation. Before each E-step, we add a trimming step, in which the less plausible observations, if the estimated model was true, are tentatively trimmed. Moreover, along the M-step, constraints on the scatter matrices are imposed, to avoid singularities and reduce the occurrence of spurious maximizers. The advantages of the new approach are shown through applications on different fields, also in comparison to recent contributions in the literature, like mixtures of skew distributions with heavier than normal tails.

CO0351: **Trimming in probabilistic clustering**
Presenter: **Gunter Ritter**, University of Passau, Germany

The normal mixture model is a popular tool for decomposing grouped multivariate data sets in their clusters. The preferred method for estimating its parameters is nowadays the likelihood paradigm. It is well known that an ML estimator does not exist but the likelihood function possesses a consistent local maximum. Consistency of a constrained MLE is due to Hathaway. A combination of both theorems leads to a trade-off between likelihood and scale balance and the SBF plot (scale balance vs. fit) leading to possible solutions. It is well known that the likelihood estimate is not robust against outliers. Some common methods of robustification, such as applying Huber's robust M-estimators to parameter estimation, adding an additional component in order to take account of outliers, or using elliptical distributions, show a clear gain in robustness. However, it was noted that they are not effective against *gross* outliers, their asymptotic breakdown point being zero. Effective protection against outliers is trimming. It leads to a transportation problem which can here be efficiently solved. The asymptotic breakdown point of the resulting method is strictly positive, an indication of robustness even against gross outliers. Some applications to synthetic and real data illustrate the method.

CO0567: **Diagnostics in finite mixture models and model-based clustering**
Presenter: **Ranjan Maitra**, Iowa State University, United States

Model-based clustering offers a principled approach to the problem of partitioning data into different groups. Data quality is important as is our confidence in the acquired clustering. We investigate the use of and develop regression-style diagnostics to quantify uncertainty in the groupings of observations as well as to identify influential and outlying observations with a view to improving model-fitting and inference.

SMALL AREA ESTIMATION

CO004 **Room Sala 3** **Chair: Domingo Morales**

CO0275: **Poverty mapping in small areas under a two-fold nested error regression model**
Presenter: **Isabel Molina**, Universidad Carlos III de Madrid, Spain
Co-authors: Domingo Morales, Yolanda Marhuenda

When the target population is naturally divided in subpopulations at two nested aggregation levels (e.g. in provinces and counties within provinces), or when the sampling design has two stages as is usual in many household surveys, it is reasonable to assume a two-fold nested error model including random effects at the two levels of aggregation, domains and subdomains. A previous empirical best method for poverty mapping is extended to a two-fold model of this kind, when the target parameters are separable. We provide analytical expressions for the EB estimators of poverty incidences and gaps obtained under the two-fold model and also a Monte Carlo algorithm for approximation of EB estimators of more complex domain or subdomain parameters. The obtained EB estimates of subdomain parameters have the good property of being consistent with the corresponding domain estimate. We provide a bootstrap estimator of the mean squared error (MSE) of EB estimators. In simulations, we compare the EB estimators of poverty incidence and poverty gap obtained under the two-fold model with the EB estimators obtained by considering a model with only domain effects or only subdomain effects, when all subdomains are sampled or when there are unsampled subdomains. Results are applied to poverty mapping in counties of the Spanish region of Valencia by gender.

CO0279: **Multivariate area level models for small area estimation**
Presenter: **Domingo Morales**, University Miguel Hernandez of Elche, Spain
Co-authors: Roberto Benavent

Multivariate area level models for estimating small area indicators are introduced. Among the available procedures for fitting linear mixed models, the residual maximum likelihood (REML) is employed. The empirical best predictor (EBLUP) of the vector of area means is derived. An approximation to the matrix of mean squared crossed prediction errors (MSE) is given and four MSE estimators are proposed. The first MSE estimator is a plug-in version of the MSE approximation. The remaining MSE estimators combine parametric bootstrap with the analytic terms of the MSE approximation. Several simulation experiments are performed in order to assess the behavior of the multivariate EBLUP and for comparing the MSE estimators. The developed methodology and software are applied to data from the 2005 and 2006 Spanish living condition surveys. The target of the application is the estimation of poverty proportions and gaps at province level.

CO0314: **Estimating poverty indicators under area-level Poisson mixed models with SAR(1) domain effects**
Presenter: **Miguel Boubeta**, Universidade da Coruna, Spain
Co-authors: Maria Jose Lombardia, Domingo Morales

Poisson mixed models are useful tools for estimating poverty indicators in territorial units, especially when there is a high degree of disaggregation. The number of persons under the poverty line in Galicia (northwest of Spain) is analysed by using area-level Poisson mixed models with SAR(1) domain effects. These models allow a structure of dependence between neighboring domains. In this context, we derive the method of moments (MM) for estimating model parameters and we obtain the empirical best predictors (EBP) of poverty proportions. We compare the EBP against alternative approaches as the synthetic and the plug-in estimators. We use the mean squared error (MSE) as a precision measure of the proposed estimator and we estimate it by parametric bootstrap. Finally, we apply the developed methodology to estimate poverty indicators in Galicia at county level.

CO0305: **On multidimensional Gaussian Markov random fields and Bayesian computation**
Presenter: **Ying C MacNab**, University of British Columbia, Canada

Proposals of multivariate Gaussian Markov random field (MGMRF) models have been advanced in tandem with developments of relevant computational solutions and strategies. The symmetric and positivity conditions for a MGMRF, or a class of MGMRFs that are typically defined by full conditionals or as linear models of coregionalization, demand carefully considered parameterization for identification and related computational strategies. Some recent works on MGMRFs are reviewed, with in-depth discussions on strategies for, and challenges in, Bayesian computation. Within the context of analysis of multivariate spatial data on finite lattice in general, and in the context of Bayesian disease mapping and small area estimation in particular, we discuss MGMRFs as prior models or as data models within Bayesian hierarchical model framework. Several examples are presented to illustrate recently proposed computational solutions and unresolved challenges.

CMSTATISTICS SESSION: ADVANCES ON COMPUTATIONAL STATISTICS AND DATA ANALYSIS I

CO035 **Room Sala 2** **Chair: Erricos Kontoghiorghes**

CO0356: **Combining multiple frequencies in multivariate volatility forecasting**
Presenter: **Alessandra Amendola**, University of Salerno, Italy
Co-authors: Vincenzo Candila, Giuseppe Storti

In a multivariate volatility framework, several options are available to estimate the conditional covariance matrix of returns. Some models, like the multivariate GARCH (MGARCH) ones, rely on daily returns while others exploit the additional information provided by the analysis of intra-daily prices, like the realized covariance (RC) specifications. An additional source of uncertainty is related to the choice of the frequency at which the intradaily returns, used to construct the RC matrices, are observed. Our interest is in analyzing the impact of these two sources of uncertainty on volatility prediction. In particular, we investigate the profitability of a prediction strategy based on combining forecasts coming from different model structures that are estimated using information at various frequencies. In order to illustrate the benefits of our approach we carry out an extensive application to portfolio allocation for a panel of U.S. stocks.

CO0472: **Genetic versus controlled approximate algorithms for regression model selection**
Presenter: **Cristian Gatu**, Alexandru Ioan Cuza University of Iasi, Romania
Co-authors: Georgiana-Elena Pascaru, Erricos John Kontoghiorghes

Algorithms for regression model selection are compared in terms of execution times and quality of obtained solution. Specifically, the recently introduced heuristic algorithm (HBBA) and implemented in the R package "lmSubsets" is compared to the genetic algorithms (GA). The targeted GAs are "gaselect", "kofnGA", and "glmulti" that are also implemented as R packages. The HBAA yields solutions having relative errors with respect to the optimum that lie within a given tolerance. Thus the quality of HBAA solutions can be properly assessed. The GA obtain also solutions that are not optimal and do not provide any information as how far they are from the optimum. The aim is to compare the HBBA and GA. Specifically, the comparison (a) investigates the maximum problem size that can be tackled by the HBBA and GA within a given reasonable

computing time; (b) determines the average relative errors of the GA solutions, say τ_{GA}, when compared to the optimum solutions which are obtained by HBBA with zero tolerance; and (c) assess the execution times of the GA and the HBBA with tolerance controlled by τ_{GA}.

CO0579: **Predicting patient's response to the treatment based on high dimensional genetic data**
Presenter: **Malgorzata Bogdan**, Wroclaw University, Poland

Several relatively new approaches will be discussed for identifying important predictors in large data bases. The common denominator of these methods is the goal of controlling the fraction of false discoveries among the selected predictors. We will present results concerning predictive properties of these methods and illustrate them with the simulation study concerning the prediction of patient's response to the treatment based on his/her genetic data.

CO0177: **A characterization theorem for the least squares piecewise monotonic data fitting**
Presenter: **Ioannis Demetriou**, University of Athens, Greece

Let a sequence of n univariate observations that include random errors be given and let k be a prescribed integer. The problem of calculating the least squares data fitting subject to the condition that the first differences of the estimated values have at most $k-1$ sign changes is considered. The choice of the positions of sign changes by considering all possible combinations of positions can be of magnitude n^{k-1}, so that it is not practicable to test each one separately. A theorem is stated that decomposes the problem into least squares monotonic estimation problems (case $k = 1$) to disjoint sets of adjacent data. Besides that the theorem allows a highly efficient calculation of the piecewise monotonic estimates, it may be useful for investigating consistency properties of these estimates.

DYNAMIC MODELLING

CG056 **Room Sala 4** **Chair: Casper Albers**

CC0248: **Structure estimation for time-varying mixed graphical models in high-dimensional data**
Presenter: **Jonas Haslbeck**, University of Amsterdam, Netherlands
Co-authors: Lourens Waldorp

Dependencies in multivariate systems (graphical models) have become a popular way to abstract complex systems and gain insights into relational patterns among observed variables. For temporally evolving systems, time-varying graphical models offer additional insights as they provide information about organizational processes, information diffusion, vulnerabilities and the potential impact of interventions. In many of these situations the variables of interest do not follow the same type of distribution, for instance, one might be interested in the relations between physiological and psychological measures (continuous) and the type of prescribed drug (categorical) in a medical context. We present a novel method based on generalized covariance matrices and kernel smoothed neighborhood regression to estimate time-varying mixed graphical models in a high-dimensional setting. In addition to our theory, we present a freely available software implementation, performance benchmarks in realistic situations and an illustration of our method using a dataset from psychopathology.

CC0513: **Dynamics of networks: The mean field approach to probabilistic cellular automata on random and small-world graphs**
Presenter: **Lourens Waldorp**, University of Amsterdam, Netherlands
Co-authors: Jolanda Kossakowski

The dynamics of networks are described using one-dimensional discrete time dynamical systems theory obtained from a mean field approach to(elementary) probabilistic cellular automata (PCA). Often the mean field approach is used on a regular graph (a grid or torus) where each node has the same number of edges and the same probability of becoming active. We consider elementary PCA where each node has two states (two-letter alphabet): "active" or "inactive" (0/1). We then use the mean field approach to describe the dynamics of a random graph and a small-world graph. The mean field can now be viewed as a weighted average of the behaviour of the nodes in the graph, since the behaviour of the nodes is determined by a different number of edges. The mean field predicts (pitchfork) bifurcations. The application we have in mind is that of psychopathology. A mental disorder can be viewed as a network of symptoms, each symptom influencing other symptoms. For instance, lack of sleep during the night could lead to poor concentration during the day, which in turn could lead to lack of sleep again by worrying that your job may be on the line. The symptom graph is more likely to be a small-world than a grid. The mean field approach then allows possible explanations of "jumping" behaviour in depression, for instance.

CC0272: **Bayesian VAR-modeling: Unraveling emotion dynamics in multivariate, multisubject time series**
Presenter: **Casper Albers**, University of Groningen, Netherlands
Co-authors: Tanja Krone, Marieke Timmerman, Peter Kuppens

Emotion dynamic research typically aims at revealing distinct information on affective functioning and regulation. Herewith, one distinguishes various elementary emotion dynamic features, which are studied using intensive longitudinal data. Typically, each emotion dynamic feature is quantified separately, which hampers the study of relationships between various features. In emotion research, the length of the observed time series is limited, and often suffers from a high percentage of missing values. We propose a Bayesian vector autoregressive model (VAR) that is useful for emotion dynamic research. The model encompasses the six central emotion dynamic features at once, and can be applied with relatively short time series, including missing data. The individual emotion dynamic features covered are: long and short term variability, granularity, inertia, cross-lag correlation and the intensity. The model can be applied to both univariate and multivariate time series, allowing to model the relationships between emotions. Further, it may model multiple individuals jointly as well as external variables and non-Gaussian observed data, and can deal with missing data. We illustrate the usefulness of the model with an empirical example using relatively short time series of three emotions, with missing time points within the series, measured for three individuals.

CC0497: **Bayesian prediction based on profile-reference data**
Presenter: **Jinfang Wang**, Chiba University, Japan
Co-authors: Shigetoshi Hosaka

The problem of predicting the future outcome for a specific subject is considered based on data resulting from a longitudinal study. This data will be referred to as the profile data, which typically contain the time-dependent responses for each subject, as well as many other variables concerning the background information on each subject. In addition to the profile data, we assume that there are also available a large reference data set containing the same time-dependent responses as in the profile data. The reference data may be obtained from national survey offices, which publish many kinds of survey data, such as data on health care. The reference data differ from the profile data in that individual data in the reference data are usually grouped and only very limited background information (e.g. gender and age) are available. We assume that the two data sets partially share a common latent structure. We propose a Bayesian model for predicting the future outcome for a specific subject by combining the profile and the reference data. We illustrate this methodology by applying a dynamic linear mixed model to predict the blood sugar level at a specific age.

ASYMPTOTIC THEORY

CG040 **Room Sala 7** **Chair: Maria Dolores Jimenez-Gamero**

CC0443: **Asymptotic confidence bands in the Spektor-Lord-Willis problem**
Presenter: **Zbigniew Szkutnik**, AGH University of Science and Technology, Poland
Co-authors: Bogdan Cmiel, Jakub Wojdyla

Confidence bands for a density function of directly observed i.i.d. data have been proposed since 1973. In the last decade, asymptotic nonparametric confidence bands have been constructed in some inverse problems, like density deconvolution, inverse regression with a convolution operator and regression with errors in variables. There seems to have been, however, no such construction for practically important inverse problems of stereology. This gap will be partially filled by constructing a kernel-type estimator for the density of balls radii in the stereological Spektor-Lord-Willis (SLW) problem, along with corresponding asymptotic uniform confidence bands and an automatic bandwidth selection method. Recall that the SLW problem consists in unfolding the distribution of random radii of balls randomly placed in an opaque medium and only observed as line segments on a random line section through the medium (like drilling through a rock or a muscle biopsy). The problem will be formulated as an ill-posed Poisson inverse problem and the construction of asymptotic confidence bands will be based, as in earlier contributions, on the supremum of a stationary Gaussian process. The finite-sample performance of the new procedures will be demonstrated in a simulation experiment.

CC0459: **Validation of positive expectation dependence**
Presenter: **Bogdan Cmiel**, Polish Academy of Sciences, Poland
Co-authors: Teresa Ledwina

There is quickly growing evidence that dependence structure of observed random vectors can not be neglected in a reliable data analysis. Extensive work of practitioners has revealed that some well-known notions as for example the correlation coefficient are not sufficient to explain complex character of many relations while some other existing or new notions can be much more useful in nowadays practice and some specific applications. Last years increasing role of the notion of positive expectation dependence has been observed in different research areas. Tests are developed for such type of dependence. The solutions are weighted Kolmogorov-Smirnov type statistics. They originate from the function valued monotonic dependence function, describing local changes of the strength of the dependence. Therefore, the inference can be supported by a simple and insightful graphical device. An asymptotic and simulation results for such tests are presented. It is shown that an inference relying on $p-$values and wild bootstrap allows to overcome inherent difficulties of this testing problem. Some simulations show that the new tests perform well in finite samples.

CC0476: **Some limit theorems in a two-sex branching model**
Presenter: **Alfonso Ramos**, University of Extremadura, Spain
Co-authors: Manuel Molina, Manuel Mota

Branching models are considered appropriate mathematical tools to describe the probabilistic evolution of dynamical systems whose components after certain life period reproduce and die, in such a way that transition from one to other state of the system is made according to a certain probability distribution. We mathematically model the demographic dynamics of biological populations with sexual reproduction where mating and reproduction can be influenced by the current number of couples in the population. We study some limit theorems in the class of two-sex models where the mating and the reproduction are affected by the number of females and males in the population.

CC0486: **Improved ϕ-divergence test statistics based on minimum ϕ^*-divergence estimator for GLIMs of binary data**
Presenter: **Nobuhiro Taneichi**, Kagoshima University, Japan
Co-authors: Yuri Sekiya, Jun Toyama

Generalized linear models of binary data including a logistic regression model and a probit model are considered. For testing the null hypothesis that the considered model is correct, ϕ-divergence family of goodness-of-fit test statistics $C_{\phi\phi^*}$ which is based on minimum ϕ^*-divergence estimator is considered. Family of statistics $C_{\phi\phi^*}$ includes a power divergence family of statistics $R^{a,b}$ which is based on minimum power divergence estimator. The derivation of an expression of continuous term of asymptotic expansion for the distribution of $C_{\phi\phi^*}$ under the null hypothesis is shown. Using the expression, a transformed $C_{\phi\phi^*}$ statistic that improves the speed of convergence to the chi-square limiting distribution of $C_{\phi\phi^*}$ is obtained. In the case of $R^{a,b}$, it is numerically shown that the transformed statistics perform much better than the original statistics and it is also numerically shown that power of the transformed statistics is almost the same as that of the original statistics.

NONPARAMETRIC REGRESSION

CG024 **Room Sala 5** **Chair: Giles Hooker**

CC0499: **Towards the development of arc length regression**
Presenter: **Theodor Loots**, University of Pretoria, South Africa
Co-authors: Andriette Bekker

The coefficients obtained from using ordinary linear regression may be severely biased with corresponding estimates lacking accuracy when the assumptions of normality are not satisfied. A new framework is proposed where the arc lengths of the kernel density functions cast over the dependent and independent variables are matched in order to yield coefficient estimates. The significance of these estimates are evaluated using resampling techniques, and model selection performed by using entropy based measures such as the Bhattacharyya divergence measure and minimum description length principle (MDL).

CC0350: **Shape constrained regression in Sobolev spaces and tests of isotonicity**
Presenter: **Michal Pesta**, Charles University in Prague, Czech Republic

A class of nonparametric regression estimators based on penalized least squares over the sets of sufficiently smooth functions is elaborated. We impose additional shape constraint - isotonia - on the estimated regression curve and its derivatives. The problem of searching for the best fitting function in an infinite dimensional space is transformed into a finite dimensional optimization problem making this approach computationally feasible. The form and properties of the regression estimator in the Sobolev space are investigated. Tests of isotonicity based on U-statistics and bootstrap are provided. An application to option pricing is presented. The behavior of the estimator is improved by implementing an approximation of a covariance structure for the observed intraday option prices.

CC0302: **Semiparametric probit model for high-dimensional clustered data**
Presenter: **Daniel Raguindin**, University of the philippines, Philippines
Co-authors: Erniel Barrios, Joseph Ryan Lansangan

A semiparametric probit model is proposed for high dimensional clustered data. The model allows flexibility in the structure to account for lost information in the process of dimension reduction. Principal components are postulated to have nonparametric effect on the dichotomous response, mitigating the lost information due to the selection of just few principal components. On the other hand, the parametric part takes advantage of inherent homogeneity within clusters, hence, a constant random intercept term accounts for data clustering. Simulation studies illustrate the advantages of the proposed model over the ordinary probit model in low dimensional cases. It also provides high predictive ability in high dimensional cases especially when the distribution of the response to the two categories is balance even in the presence of misspecification error.

CC0524: **Two-step estimation for varying coefficient regression models with censored data**
Presenter: **Seong Jun Yang**, Hankuk University of Foreign Studies, Korea, South
Co-authors: Cedric Heuchenne, Ingrid Van Keilegom

Estimators of the coefficient functions for the varying coefficient model are proposed where the response is subject to random right censoring.

The model includes different coefficient functions depending on various covariates. Since multivariate smoothing is unavoidable under the model, smooth backfitting is applied to avoid “the curse of dimensionality". The estimation method is based on the Bucklely-James type transformation, where the estimators achieved by Koul-Susarla-Van Ryzin type transformation are used for primary estimators of the coefficient functions. Asymptotic normality of the proposed estimators are given, and numerical studies are shown to illustrate the reliability of the estimators.

Friday 26.08.2016 11:00 - 12:05 Parallel Session O – COMPSTAT

ADVANCES AND NEW METHODOLOGIES IN LIFETIME DATA ANALYSIS: SURVIVAL AND RELIABILITY

CO049 **Room Sala 1** **Chair: Juan Eloy Ruiz-Castro**

CO0308: **Survival analysis for semi-Markov processes**
Presenter: **Vlad Barbu**, Universite de Rouen, France

Semi-Markov processes and Markov renewal processes represent a class of stochastic processes that generalize Markov and renewal processes. As it is well known, for a discrete-time (respectively continuous-time) Markov process, the sojourn time in each state is geometrically (respectively exponentially) distributed. In the semi-Markov case, the sojourn time distribution can be any distribution on N (respectively on R). This is the reason why the semi-Markov approach is much more suitable for applications than the Markov one. The purpose is to investigate some survival analysis and reliability problems for semi-Markov processes and to address some statistical topics. We start by briefly introducing the discrete-time semi-Markov framework, giving some basic definitions and results. These results are applied in order to obtain closed forms for some survival or reliability indicators, like survival/reliability function, availability, mean hitting times, etc. The nonparametric estimation of the characteristics of a semi-Markov system and of the associated survival/reliability indicators is considered. A particular attention is given to censored data in semi-Markov framework.

CO0360: **Analysis of software bug data across version releases with application to optimal version release**
Presenter: **Simon Wilson**, Trinity College Dublin, Ireland
Co-authors: Sean ORiordain

An emerging trend in software over the last 5 years has been to maintain a schedule of frequent version releases in order to remain competitive and pass on important security updates to users as quickly as possible. While software testing is still done, software users are encouraged to report bugs to the developer. We look at some simple models for the times to bug discovery across successive versions of software. We propose various ways in which the dependence between releases can be modelled. We fit the model to data on bug discovery for the Mozilla Firefox browser. Finally we use the fitted model to suggest an optimal time between releases that trades off the costs of releasing faulty software with the opportunity costs of delaying release.

CC0249: **The Dagum regression model in survival analysis**
Presenter: **Mariangela Zenga**, Milano-Bicocca University, Italy
Co-authors: Juan Eloy Ruiz-Castro, Filippo Domma

The Dagum distribution is a Burr III distribution with an additional scale parameter. It is closely related to the Burr XII distribution and, more generally, turns out to be a special case of the Generalized Beta distribution. Even if the Dagum model has been used in studies of income and wage distribution as well as wealth distribution, only recently it was introduced in the field of the survival analysis and the reliability. The hazard rate of this model is very flexible; in fact, it is proved that, according to the values of the parameters, the hazard rate of the Dagum distribution has a decreasing, or a Upside-down Bathtub, or Bathtub and then Upside-down Bathtub failure rate. Moreover some features of this distribution (as the reversed hazard rate, the mean and variance of the random variables residual life and reversed residual life and their monotonicity properties) were studied. We will consider the observed heterogeneity depending on covariates in a regression model with a Dagum distribution.

INTERVAL AND DISTRIBUTIONAL DATA IN DATA SCIENCE

CO037 **Room Sala 3** **Chair: Javier Arroyo**

CO0345: **Joint similarity measures defined with quasi-arithmetic means**
Presenter: **Etienne Cuvelier**, ICHEC - Brussel, Belgium

A lot of data analysis methods are based on similarity or dissimilarity measures but, most of the times, these measures are defined for one type of data (real multidimensional data, interval data, functional data, ...). This fact implies that all the techniques of knowledge extraction based on such measures can be performed only on the data type for which they are defined. But the description and the modelling of real situations require the joint use of several kind of data. We propose a new technique of combination of different measures in one single result. The method is based on Quasi-Arithmetics Means using Archimedean Generators. Quasi-Arithmetic Means with this kind of generators have several advantages to compute a resulting measure starting from several measures (computed on different types of data describing the same concept or individual): they allow to choose to emphasize the similarity or the dissimilarity between objects, they have flexible parameters,its possible to mix similarities and dissimilarities to compute a resulting similarity (or dissimilarity). The resulting measure (similarity or dissimilarity) can be used in any existing algorithm based on such measures: clustering, supervised classification and so on. We will give some examples of use of this method data mixing functional data and other types of data.

CO0317: **HistDAWass package: An R tool for forecasting histogram-valued data**
Presenter: **Javier Arroyo**, Universidad Complutense de Madrid, Spain
Co-authors: Antonio Irpino

In the framework of Symbolic Data Analysis, a relatively new approach to the statistical analysis of multi-valued data, we consider histogram-valued data, i.e. where each statistical unit is described by several univariate histograms. This approach is appropriate to analyze statistical units that aggregate a set of values. We present the HistDAWass package for R, which includes statistical methods to analyze this kind of data. The methods and the basic statistics for histogram-valued data are mainly based on the L_2 Wasserstein metric between distributions, i.e. a Euclidean metric between quantile functions. The package contains unsupervised classification techniques, least squared regression and tools for histogram-valued data and for histogram time series. We will show the main features of the package and applications to histogram-valued datasets extracted from different real-life contexts.

CC0492: **Improved scale-space analysis for interval-valued data**
Presenter: **Kee-Hoon Kang**, Hankuk University of Foreign Studies, Korea, South
Co-authors: Cheolwoo Park, Yongho Jeon

With the rapid advancement of computing technology and storage capacity, both the size of data and the complexity of their structure have significantly increased. These enormous data are sometimes converted into new types of data such as intervals, histogram, and trees, etc. Interval-valued data represent uncertainty or variability and offer richer and more complex information on trend and variation of the underlying structure than single-valued data. One can sample single-valued data from interval-valued data by assuming a uniform distribution. This can be improved if the single-valued data are generated from the actual underlying distribution rather than a uniform distribution. Recently, a nonparametric method of estimating a joint density using the interval-valued data is developed. It treats the observed set of hyper-rectangles as a multivariate histogram that can be approximated to locally weighted Gaussian kernel functions. We consider this approach with SiZer map for interval-valued data in a nonparametric regression setting.

CMSTATISTICS SESSION: ADVANCES ON COMPUTATIONAL STATISTICS AND DATA ANALYSIS II
CO112 **Room Sala 2** **Chair: Ori Davidov**

CC0529: **A robust fuzzy clustering method for non-precise data based on trimmed regions**
Presenter: **Ana Belen Ramos-Guajardo**, University of Oviedo, Spain
Co-authors: Maria Brigida Ferraro

The use of fuzziness in data analysis to capture the imprecision inherent in several sources of information has been deeply addressed in recent works. Additionally, various methods for clustering fuzzy data, including also fuzziness in the clustering process, have been developed in the last decades. Most of fuzzy clustering methods based on a $k-$means procedure do not take into account the influence of outliers in the data set. As an attempt to deal with this problem, a robust fuzzy clustering method is proposed based on trimming techniques, which have been shown to be very intuitive and applicable in many general spaces. The proposed methodology considers the so-called trimmed fuzzy mean and its natural empirical estimator. The trimmed fuzzy mean is defined on the basis of a generalized distance as the value minimizing the variance of a random fuzzy set (or fuzzy random variable) over the possible trimmed regions. Finally, the behaviour of the method is empirically analyzed by means of some simulation studies and its applicability is shown in a real-life example.

CO0403: **Non-reduced versus reduced-bias estimators of the extreme value index: Efficiency and robustness**
Presenter: **Ivette Gomes**, FFCUL, Universidade de Lisboa and CEAUL, Portugal

The *extreme value index* (EVI) is the primary parameter of extreme events. The EVI is used to characterize the tail behavior of a distribution, and it helps to indicate the size and frequency of certain extreme events under a given probability model: for large events, the bigger the EVI is, the heavier is the right-tail of the underlying parent distribution. The Lehmer mean of order p of the k log-excesses over the $k+1$-th upper order statistic has been recently considered in the literature for the estimation of a positive EVI, associated with large extreme events. Such a Lehmer mean of order p generalizes the arithmetic mean $(p = 1)$, the classical Hill estimator of a positive EVI, and for $p > 1$ has revealed to be very competitive for small values of the EVI, comparing favorably with one the simplest classes of reduced-bias EVI-estimators, a corrected-Hill estimator. Now, the comparison to other EVI-estimators is performed, and some information on the robustness of such a general class is provided, including its resistance to possible contamination by outliers.

CO0164: **Univariate analysis of compositional data using weighted balances**
Presenter: **Karel Hron**, Palacky University, Czech Republic
Co-authors: Peter Filzmoser, Alzbeta Gardlo

Compositional data, observations carrying exclusively relative information (with units like percentages, mg/kg, mg/l, etc.), have specific properties that are not compatible with the Euclidean geometry requirement of most standard statistical methods. In order to represent compositional data in the usual Euclidean geometry, they need to be expressed in orthonormal coordinates prior to statistical processing. As it is not possible to construct standard Cartesian coordinates for compositions that assign a coordinate for each of the parts separately, a choice of interpretable orthonormal coordinates is of particular interest. Although recent experiences show clear advantages of such coordinates where the first coordinate aggregates information from log-ratios for a particular compositional part of interest, their usefulness is limited if there are distortions like rounding errors or other data "problems" in the involved parts. The purpose is to introduce a "robust" (weighted) version of these coordinates, called weighted balances, where the remaining parts (with respect to the part of interest) in the first coordinate are weighted in a way that is relevant to the aims of the statistical analysis. Such weights can be, e.g., derived according to quality assessment analysis and elements of classical/robust variation matrix of compositions. Methodological outputs are accompanied by a real-world example from metabolomics.

SEMIPARAMETRIC REGRESSION
CG007 **Room Sala 5** **Chair: Fernando Quintana**

CC0233: **A semiparametric model for generalized Pareto regressions based on a dimension reduction assumption**
Presenter: **Julien Hambuckers**, University of Goettingen, Germany
Co-authors: Cedric Heuchenne, Olivier Lopez

A regression model is considered in which the tail of the conditional distribution of the response can be approximated by a Generalized Pareto distribution. Our model is based on a semiparametric single-index assumption on the conditional tail index $\gamma(x)$; while no further assumption on the conditional scale parameter is made. The underlying dimension reduction assumption allows the procedure to be of prime interest in the case where the dimension of the covariates is high, in which case the purely nonparametric techniques fail while the purely parametric ones are too rough to correctly fit to the data. We derive asymptotic normality of the estimators that we define, and propose an iterative algorithm in order to perform their practical implementation. Our results are supported by some simulations. To illustrate the proposed approach, the method is applied to a new database of operational losses from the bank UniCredit.

CC0211: **Switching meta-regression model in high dimensional data**
Presenter: **Ivy Corazon Ancog**, Bohol Island State University and University of the Philippines, Philippines
Co-authors: Erniel Barrios, Joseph Ryan Lansangan

The aim is to estimate a semiparametric mixed switching meta-regression model with high dimensional predictors in a hybrid of cubic smoothing splines (fixed effect component) and a Restricted Maximum Likelihood (REML) (random effect component) embedded in the backfitting algorithm. Sparse Principal Component Analysis is used in dimension reduction and variable selection but a nonparametric function of the sparse principal components is postulated in the model to address the decline in predictive ability of the model due to the use of only few components of the high dimensional predictors. We also consider regime switch (two possible groupings) identified via Support Vector Machine Classifier. Simulation study shows that the proposed procedure yields better predictive ability (Mean Absolute Percentage Error) than the Ordinary Least Square (OLS) method.

CC0198: **Illumination problems in digital images: A statistical point of view**
Presenter: **Segolen Geffray**, Universite de Strasbourg, France
Co-authors: Nicolas Klutchnikoff, Myriam Vimond

Interest is focused on a multi-dimensional signal R observed on a grid in the presence of both an illumination artifact and an additional additive bounded variance centered noise ε. The illumination artifact consists of colour or grey level intensity variations which are seen on the sampled image but which are not present in R. Such an assumption is classically modelled using a function L which acts multiplicatively on R. Our goal is to estimate R from observations of a random variable Y which obeys the regression model $Y = RL + \varepsilon$. We identify and propose a solution to an identifiability issue. We construct a consistent multi-step estimation procedure. We first make use of any consistent denoising method to estimate from the noisy data the gradient of the logarithm of the regression function. We assume that the artefact L consists of "smooth" variations. Then we project the denoised auxiliary estimate on a finite basis of "smooth" functions. We deduce the final estimator of R so that the proposed identifiability constraint is satisfied. An additional Monte Carlo computation is used to approximate a relevant integral. We derive an upper bound for the sup-norm risk of our estimator. Applications to different images are presented.

POSTER SESSION III

CP001 **Room Ground Hall** **Chair: Francisco Torres-Ruiz**

CP0457: **Supervised classification using a distance-depth function**
Presenter: **Itziar Irigoien**, University Basque Country, Spain
Co-authors: Concepcion Arenas, Francesc Mestres

Supervised classification is used by researchers in a wide variety of fields as in taxonomic classification; in morphometric analysis for species identification; in ecological problems addressed to test the presence or absence of a particular species; in marine ecology to evaluate the similarity of distinct populations and to classify units of unknown origin to known populations; in genetic studies in order to summarize the genetic differentiation between groups or in the biomedical context, predicting the diagnostic category of a sample on the basis of its gene expression profile and some clinical features. A novel classifier rule is introduced based on an improvement of the distance-based discriminant (DB-discriminant), taking into account a depth function. This new model combines the DB-rule and the maximal depth classifier, obtaining a classifier that is often more accurate than both methods separately. To demonstrate its effectiveness the new classifier was compared with the DB-rule and the $k-$nearest neighbor classification method, using high-dimensional class-imbalanced cancer data sets, and evaluating the leave-one-out error rate, the generalized correlation coefficient, the sensitivity, the specificity and the positive predicted value for each class. The results show the good performance of the new classifier.

CP0467: **Estimation in the functional convolution model**
Presenter: **Tito Manrique**, UMR MISTEA - INRA Montpellier SUPAGRO, France
Co-authors: Christophe Crambes, Nadine Hilgert

An estimator is proposed for the unknown function in the Functional Convolution Model, which studies the relationship between a functional covariate $X(t)$ and a functional response $Y(t)$ through the following equation $Y(t) = \int_0^t \theta(s)X(t-s)ds + \varepsilon(t)$, where θ is the function to be estimated and ε is an additional functional noise. In this way we can study the influence of the history of X on $Y(t)$. We use the Continuous Fourier Transform to define an estimator of θ. The transformation of the convolution model results in the Functional Concurrent Model associated, in the frequency domain, namely $\mathcal{Y}(\xi) = \beta(\xi)\mathcal{X}(\xi) + \varepsilon(\xi)$. In order to estimate the unknown function β, we extended the classical ridge regression method to the functional data framework. We establish consistency properties of the proposed estimators and illustrate our results with some simulations.

CP0505: **A novel two-step iterative approach for clustering functional data**
Presenter: **Zuzana Rostakova**, Slovak Academy of Sciences, Slovakia
Co-authors: Roman Rosipal

An important task in functional data analysis is to divide a dataset into subgroups with similar profiles, or clusters. We address a problem in which classical functional data clustering techniques may fail when curve misalignment is present. Solutions in which registration or temporal alignment of the whole dataset precede the clustering step result in rapid distortions in the curve shapes when a dataset consists of many different curve profiles. Methods developed for simultaneous registration and clustering of curves mainly deal with linear transformation of time. This solution may also lead to unsatisfactory alignment when profiles of the curves are complex or the source of misalignment has a nonlinear character. We propose and validate a novel two-step approach, which iteratively combines clustering using a modified Dynamic Time Warping algorithm with the registration step applied separately to curves within estimated clusters. On generated and real functional data representing the sleep process we demonstrate the validity of the approach by measuring improvement in similarity between aligned curves in comparisons to: a) the case when clustering and registration steps are applied separately and b) other methods (e.g. k-means alignment, joined probabilistic curve clustering and alignment) for simultaneous curve registration and clustering.

CP0510: **Prediction of disease risk by high-dimensional genetic and environmental data**
Presenter: **Norbert Krautenbacher**, Technical University of Munich and Helmholtz Center Munich, Germany
Co-authors: Christiane Fuchs, Fabian Theis

The aim is to investigate the situation of having high-dimensional genetic and environmental data of individuals where the goal is to build a prediction model for the risk of suffering from the disease asthma. At the study one was also interested in the influence of the specific exposure variable farm-environment, so that a sample of the population should contain an appropriate number of observations with the combination farm/asthma. Since in the population both categories occur only rarely, a simple random sample would require a big sample size. In practice, however, it is not possible to take such a big sample, since collecting genomic data in terms of hundreds of thousands to millions of single-nucleotid polymorphisms (SNPs) is cost-intensive. Thus, a stratified random sample was taken from the population. Therefore, for analyzing the final sample two main issues occur: first, one has to correct for the arisen sample selection bias when learning and evaluating on biased training and test data sets. Second, the present genetic data containing 2.5 million SNPs have to be incorporated as features for dimension reduction and feature selection techniques which require special solutions.

CP0504: **Modified profile likelihood in complex models with many nuisance parameters**
Presenter: **Claudia Di Caterina**, University of Padova, Italy
Co-authors: Nicola Sartori

It is well known that usual frequentist inference procedures for a parameter of interest are generally highly inaccurate when dealing with statistical models where the number of nuisance parameters is large relative to the sample size. Among the alternative proposals put forward in the literature, the modified profile likelihood has proved to represent a valid solution to the problem. Specifically, the approximation to such pseudo-likelihood previously introduced allows us to overcome some difficulties related with its computation outside the class of exponential and group family models. Nevertheless, even this modification of the profile likelihood can be hard to obtain analytically under moderately complex scenarios. In order to further enlarge the domain of applicability of this technique, Monte Carlo simulation can be used to evaluate some expected values involved in the modified profile likelihood. It is shown how such an approach succeeds in providing a reliable inference on the parameter of interest in various frameworks, all considering a panel data structure: microeconometric fixed effects models with continuous or discrete response, models for datasets with missing values in the dependent variable or in the covariates, and parametric survival models for censored data.

CP0217: **Flexible Birnbaum-Saunders models**
Presenter: **Heleno Bolfarine**, University of Sao Paulo, Brazil

We introduce a new extension of the Birnbaum-Saunders distribution as a follow up to the family of skew-flexible-normal distributions. This extension produces a family of Birnbaum-Saunders distributions including densities that can be unimodal as well as bimodal. This flexibility is important in dealing with positive bimodal data, given the difficulties experienced by the use of mixtures of distributions when bimodality is present. Some basic properties of the new distribution are studied including moments. Parameter estimation is approached via the method of moments and also by maximum likelihood, including a derivation of the Fisher information matrix. Computational aspects of maximum likelihood implementation is discussed. Real data illustrations indicate satisfactory performance of the new model.

CP0235: **Influence of missing data on the estimation of the number of components of a PLS regression**
Presenter: **Frederic Bertrand**, Universite de Strasbourg, France
Co-authors: Nicolas Meyer, Myriam Maumy-Bertrand

Partial Least Squares regression (PLSR) is a multivariate model for which two algorithms (SIMPLS or NIPALS) can be used to provide its parameters estimates. The NIPALS algorithm has the interesting property of being able to provide estimates with incomplete data and this has been extensively studied in the case of principal component analysis for which the NIPALS algorithm has been originally devised. Nevertheless, the literature gives no clear hints at the amount and patterns of missing values that can be handled by this algorithm in PLSR and to what extent the model parameters estimates are reliable. We study the NIPALS behavior, when used to fit PLSR models, for various proportions and pattern of missing data (at random or completely at random). Comparisons with multiple imputation are done. The NIPALS algorithm tolerance to incomplete data sets depends on the sample size, the proportion of missing data and the chosen component selection method and a proportion of 30% of missing data can be given as an empirical maximum for a reliable components number estimation. Above this value, whatever the criterion considered, except the Q_2, the number of components in PLSR is far from the true one and may hence give misleading conclusions.

CP0415: **A visualized measure vector of departure from double symmetry for square contingency tables**
Presenter: **Shuji Ando**, Novartis Pharma KK, Japan
Co-authors: Kouji Tahata, Sadao Tomizawa

For square contingency tables, models having a structure of double symmetry have been proposed. The double symmetry indicates that both structure of symmetry and point symmetry hold. For measuring the degree of departure from model having structure of double symmetry, we propose a two-dimensional measure vector that can simultaneously measure the degree of departure from symmetry and the degree of departure from point symmetry. Using the proposed measure vector, we can see the degree of departure from double symmetry, while distinguishing the degree of departure from symmetry and the degree of departure from point symmetry.

CP0536: **Model-based ordination method for overdispersed count data**
Presenter: **Jenni Niku**, University of Jyvaskyla, Finland
Co-authors: Francis Hui, Sara Taskinen, David Warton

Unconstrained ordination methods are commonly used in ecology to visualize the relationships between different sites in terms of their species composition. Classical unconstrained ordination methods, such as non-metric multidimensional scaling, are algorithm-based techniques, which are developed and implemented without directly taking into account the statistical properties of the multivariate data. The ignorance of important data properties (such as mean-variance relationship) can then yield to misleading results. We consider a model-based approach to unconstrained ordination. The method uses a generalized linear latent variable model to produce an ordination plot. As usual, a model-based approach gives us tools for diagnostics, model selection and statistical inference. Examples and simulations are presented to illustrate the method in case of overdispersed count data. The results are compared with the results based on classical unconstrained ordination methods.

CP0478: **Using simulated annealing and variable neighborhood search procedures for estimating the Hubbert diffusion process**
Presenter: **Francisco Torres-Ruiz**, Granada, Spain
Co-authors: Istoni da Luz-Sant-Ana, Patricia Roman-Roman

A problem of great current interest is how to accurately chart the progress of oil production. It is well known that oil exploration is cyclical and that after the oil production reaches its peak in a specific system, a fatal decline will begin. In this context, M.H. Hubbert developed his peak theory in 1956, based on a bell-shaped curve that bears his name. We consider a stochastic model, based on the theory of diffusion processes, associated with the Hubbert curve. The problem of the maximum likelihood estimation of the parameters for this process is also considered. Since a complex system of equation appears, whose solution cannot be guaranteed via the classical numerical procedures, we suggest the use of metaheuristic optimization algorithms such as Simulated Annealing and Variable Neighborhood Search. Some strategies are suggested for bounding the space of solutions, and a description is provided for the application of the algorithms selected. In the case of the Variable Neighborhood Search algorithm, a hybrid method is proposed in which it is combined with Simulated Annealing. In order to validate the theory developed here, we carry out some studies based on simulated data and consider some real scenarios from crude oil production data in Norway and Kazakhstan.

CP0236: **New insights in Approximate Bayesian Computation algorithms for network reverse-engineering**
Presenter: **Myriam Maumy-Bertrand**, Universite de Strasbourg, France
Co-authors: Nicolas Jung, Frederic Bertrand, Khadija Musayeva

Elucidating gene regulatory network is an important step towards understanding the normal cell physiology and complex pathological phenotype. Reverse-engineering consists in using gene expression over time or over different experimental conditions to discover the structure of the gene network in a targeted cellular process. The fact that gene expression data are usually noisy, highly correlated, and have high dimensionality explains the need for specic statistical methods to reverse engineer the underlying network. Among known methods, Approximate Bayesian Computation (ABC) algorithms have not been thoroughly studied for network inference. Due to the computational overhead their application is also limited to a small number of genes. We have developed a new multi-level ABC approach that has less computational cost. At the first level, the method captures the global properties of the network, such as scale-freeness and clustering coecients, whereas the second level is targeted to capture local properties, including the probability of each couple of genes being linked. Our approach is evaluated on longitudinal expression data in Escherichia coli.

CP0264: **Runway excursions: Risk assessment with Bayesian network models**
Presenter: **Fernando Calle-Alonso**, University of Extremadura, Spain
Co-authors: Carlos Javier Perez Sanchez, Eduardo Sanchez Ayra, David Rios Insua

Aviation has grown from the first commercial flight in 1914 up to the present, achieving for the first time more than 100,000 landings a day in 2014. Forecasts are very optimistic, expecting to duplicate the number of passengers in 2030. Although accidents are extremely rare in this sector, their consequences can be dramatic. One of the most important problems are runway excursions, taking place when the plane overexceeds the end (over-run) or the lateral (ver-off) of the landing strip. In order to analyze the most influential variables for runway excursions and to provide information on risk scenarios in landing procedures, a Bayesian network-based approach has been considered. A model selection has been performed among different Bayesian networks. The proposed methodology has been applied to a database containing a number of variables related to landing operations on three landing strips. A runway excursion risk ranking is provided for the three landing strips, and the main risk variables (and states) are identified.

CP0381: **Bayesian estimation of extreme value mixture models: Simplifications and enhancements**
Presenter: **Daniela Laas**, University of Saint Gallen, Switzerland

Extreme value mixture models combine the generalized Pareto distribution for the tail approximation with a parametric, semiparametric, or non-parametric model for the body. The Bayesian estimation of these models has the advantage of deriving a full distribution of thresholds between the body and tail models, but the multimodality of the posterior distribution frequently leads to convergence or mixing problems in the Markov chain Monte Carlo simulation. In addition, the parameter dependence and a commonly used simplification in the estimation of the tail fraction may lead to inefficient sampling and imprecise estimation results. A comprehensive simulation study and empirical application to historical insurance losses show that the parallel tempering algorithm can substantially enhance the convergence behavior of the Markov chains and reduce the lag-50 autocorrelations by forty per cent or more. A reparameterisation of the generalized Pareto distribution to overcome the threshold dependence of

the scale parameter does not seem to be necessary in a real world setting with limited sample sizes. Similar estimation results under the simplified sampling algorithm with maximum likelihood approximation of the tail fraction and a newly developed full Bayesian approach further support the use of the simplified method.

Friday 26.08.2016 12:15 - 13:15 Room: Sala 1 Chair: Gil Gonzalez-Rodriguez Keynote 1

COMPSTAT Keynote Talk. Random objects: Functional data in nonlinear subspaces and Frechet regression

Speaker: **Hans-Georg Mueller, University of California Davis, United States**

Random objects will be illustrated for three commonly encountered scenarios. A general characteristic of random objects is that one has an i.i.d. sample of these objects which lie in a metric space that often has additional properties. The goal is to quantify mean and variation in a sensible way. In the first scenario, functional data that lie on a smooth isometric manifold will be considered. This includes time-warped functional data, where manifold learning with Isomap is shown to provide interpretable data analysis. The second scenario concerns functional data that are density functions. A transformation to a Hilbert space, centered around the Wasserstein mean, then leads to sensible modes of variation. In a third scenario we consider random objects that belong to a more general metric space. We view these as responses in a regression model that features scalar or vector predictors.

Saturday 27.08.2016 11:00 - 11:50 Room: Sala 1 Chair: Ana Colubi Keynote 2

A bridge between high-dimensional and functional data: Functional Cox model

Speaker: **Jane-Ling Wang, University of California Davis, United States**

There is a close relation between high-dimensional data and functional data. For instance, densely observed functional data can be viewed as high-dimensional data endowed with a natural ordering. We explore the opposite question whether one can find a proper ordering of high-dimensional data so they can be reordered and viewed as functional data. Stringing is such a method that takes advantage of the high dimensionality by representing such data as discretized and noisy observations that originate from a hidden smooth stochastic process. It transforms high-dimensional data to functional data so that established techniques from functional data analysis can be applied for further statistical analysis. We illustrate the advantage of the stringing methodology through several data sets. In one of the applications, stringing leads to the development of a new Cox model that accommodates functional covariates. Theoretical properties of the proposed functional Cox model will be explored.

Sunday 28.08.2016 16:40 - 17:30 Room: Sala 1 Chair: Enea Bongiorno Keynote 3

Recent extensions of independent component analysis

Speaker: **Hannu Oja, University of Turku, Finland**

In independent component analysis (ICA) it is assumed that the observed random vectors are linear combinations of latent, mutually independent random variables called the independent components. It is then often assumed that only the non-Gaussian independent components are of interest and the Gaussian components are treated as noise. The aim is to extract the non-Gaussian components as well as to isolate the signal and noise subspaces. In this way, ICA can recover the patterns that cannot be identified by classical principal component analysis (PCA). Popular approaches to solve the problem are briefly reviewed and their extensions to tensor data, multivariate time series and functional data are discussed. The most popular and practical approaches to solve the independent component problem are projection pursuit (e.g. FastICA) and utilizing various moment based scatter matrices (e.g. FOBI and JADE). We discuss these approaches for vector valued data and their extensions to tensor (matrix) dataas well as to multivariate time series. Extensions of ICA will be discussed for the cases where the Euclidean space is replaced by a Hilbert space of functions on an interval.

Friday 26.08.2016 15:00 - 16:15 Parallel Session B – CRoNoS FDA 2016

VARIABLE SELECTION AND SPARSE MODELS FOR FDA

CO019 **Room Sala 2** **Chair: Nathalie Villa-Vialaneix**

CO0152: **Lasso estimation of local independence graphs based on Hawkes processes**

Presenter: **Vincent Rivoirard**, Paris Dauphine University, France

Functional connectivity in neuroscience is considered as one of the main features of the neural code. It is nowadays possible to obtain the spike activities of tens to hundreds of neurons simultaneously and the issue is then to infer the functional connectivity thanks to those complex data. To deal with this problem, we consider estimation of sparse local independence graphs by using models based on multivariate Hawkes processes. Such popular counting processes depend on an unknown functional parameter to be estimated by linear combinations of a fixed dictionary. To select coefficients, we propose a Lasso-type procedure, where data-driven weights of the ℓ_1-penalty are derived from Bernstein inequalities. Our tuning procedure is proven to be robust with respect to all the parameters of the problem, revealing its potential for concrete purposes.

CO0153: **Interval sparsity for functional inverse regression**

Presenter: **Remi Servien**, INRA, France

Co-authors: Victor Picheny, Nathalie Villa-Vialaneix

The focus is on the functional regression model in which a real random variable has to be predicted from functional predictors. We use the semiparametric framework of Sliced Inverse Regression (SIR). SIR is an effective method for dimension reduction of high-dimensional data which computes a linear projection of thepredictors in a low-dimensional space, without loss on regression information. We address the issue of variable selection in functional SIR in order to improve the interpretability of the components. We extend the approaches of variable selection developped for multidimensional SIR to select intervals rather than separated evaluation points in the definition domain of functional predictors. SIR is formulated in different and equivalent ways which are alternatively used to produce ridge estimates of the components which are shrinked afterwards with a group-LASSO like penalty. An iterative procedure which automatically defines the relevant intervals is also introduced to provide a priori information to the sparse model. The approach is proved efficient on simulated data.

CC0157: **Bayesian functional linear regression with sparse step functions**

Presenter: **Paul-Marie Grollemund**, University of Montpellier, France

Co-authors: Christophe Abraham, Pierre Pudlo, Meili Baragatti

Scalar-on-function regression is a common tool to explain scalar outcomes from functional predictors. It is helpful in many applications, for instance in agronomy to explain the yield of parcels from temperature curves. Our aim is to estimate the coefficient function with an interpretable estimate. The idea is to recover the relevant intervals of the support that explain the scalar outcome. For instance it is important for farmers to know the periods during which temperature plays a main role in the yield of their parcels. We will first define the notion of an interpretable estimate and present the Bayesian model. Then we will present the estimator and the numerical process to select intervals. We will also propose a way to take prior knowledge into account. Eventually the proposed method will be applied on simulated datasets and real world datasets.

SUMMER COURSE: SESSION I

CC036 **Room Sala 1** **Chair: Jane-Ling Wang**

CK0204: **Functional data Analysis: From basics to current topics of interest**

Presenter: **Jane-Ling Wang**, University of California Davis, United States

Co-authors: Hans-Georg Mueller

An introduction into the most commonly used methods of FDA. These include Functional Principal Component Analysis (FPCA) and the related concept of modes of variation, which is based on simple statistical notions such as mean and covariance function of a random process that can be inferred from the data. FPCA is an important dimension reduction tool and in sparse data situations can be used to impute functional data that are sparsely observed.Another core topic of FDA is functional regression, where one pairs functions or scalars as predictors with responses that are also functions or scalars. For the case where the predictors include functions, a difficult step that requires regularization is the inversion of a covariance operator, which is an ill-posed problem. Such an inverse problem is also related to some forms of functional correlation, which will be another core topic. Nonlinear methods have also found increasing interest. These include polynomial and quadratic regression relations, dimension reduction methods such as additive, continuously additive and index models, and other nonlinear approaches. Further topics of interest that may be covered are warping and manifold learning, the learning of time dynamics from observed realizations of the underlying stochastic process, multivariate and repeatedly observed functional data and stringing of high-dimensional data into functional data.

Friday 26.08.2016 16:45 - 18:25 Parallel Session C – CRoNoS FDA 2016

NEW PROCESSINGS FOR FUNCTIONAL DATA

CO021 **Room Sala 2** **Chair: Frederic Ferraty**

CO0151: **On the range of integration of a functional linear model**
Presenter: **Giles Hooker**, Cornell University, United States
Co-authors: Peter Hall

A conventional linear model for functional data involves expressing a scalar response variable in terms of a weighted integral of an explanatory functional covariate in which the weighting function is a parameter to be estimated. However, in some problems the support of this weight is a proper and unknown subset of the function's domain and is a quantity of particular practical interest. Motivated by a real-data example involving particulate emissions, we develop methods for estimating the upper end of this support along with other parameters in the functional linear model. We introduce techniques for selecting tuning parameters; and we explore properties of our methodology using both simulation and the real-data example mentioned above. Additionally, we derive theoretical properties of the methodology, and discuss implications of the theory. Our theoretical arguments give particular emphasis to the problem of identifiability.

CO0171: **Stable and predictive functional domain selection with application to brain images**
Presenter: **Ah Yeon Park**, University of Cambridge, United Kingdom
Co-authors: John Aston, Frederic Ferraty

Motivated by increasing trends of relating brain images to a clinical outcome interest, we propose a functional domain selection (FUDOS) method that effectively selects subregions of the brain associated with the outcome. We view each individual's brain as a 3D functional object, and the aim is to distinguish the region where $\beta(t) = 0$ from $\beta(t) \neq 0$, where t denotes voxel location. FUDOS is composed of two stages of estimation. We first segment the brain into several small parts based on the correlation structure. Then, potential subsets of the obtained segments are built and their predictive performance are evaluated to select the best subset. The estimation involves two tuning parameters, displaying a joint effect on the amount of selection. To determine the best subset, we perturb data many times and select subregions that appear in the selected subsets with high probability. Extensive simulations are conducted, and the effectiveness in selecting the true subregion is evaluated. We finally apply the proposed method to identify subregions of the brain associated with conversion to Alzheimer's Disease in patients with mild cognitive impairment. Due to the sparseness, the result can provide more interpretable information about the association. Moreover, the selected subregions show high associations with the expected anatomical brain areas known to have memory-related functions.

CO0174: **Clustering functional data using projections**
Presenter: **Aurore Delaigle**, University of Melbourne, Australia
Co-authors: Tung Pham, Peter Hall

The aim is to show that, in the functional data context, by appropriately exploiting the functional nature of the data, it is possible to cluster the observations "asymptotically perfectly". We demonstrate that this level of performance can often be achieved by the k-means algorithm as long as the data are projected on a carefully chosen finite dimensional space. We propose an iterative algorithm to choose the projection functions in a way that optimises clustering performance, and we argue that, in order to avoid peculiar solutions, we need to use a weighted least-squares criterion. We apply our iterative clustering procedure on simulated and real data, where we show that it works particularly well.

CO0165: **Nonparametric regression on contaminated functional predictor with application to hyperspectral data**
Presenter: **Frederic Ferraty**, Mathematics Institute of Toulouse, France

The focus is on scalar-on-function nonparametric regression when only a contaminated version of the functional predictor is observable at some measurement grid. To override this common setting, a kernel presmoothing step is achieved on the noisy functional predictor. Then, the kernel estimator of the regression operator is built using the smoothed functional covariate instead of the original corrupted ones. The rate of convergence is stated for this nested-kernel estimator (i.e. the smoothing parameter of the presmoothing stage minimizes some predictive criterion) with a special attention on high-dimensional setting (i.e the size of the measurement grid is much larger than the sample size). The proposed method is applied on simulated datasets in order to assess its finite-sample properties. Our methodology is further illustrated on a real hyperspectral dataset involving a supervised classification problem.

SUMMER COURSE: SESSION II

CC037 **Room Sala 1** **Chair: Jane-Ling Wang**

CK0208: **Functional data Analysis: From basics to current topics of interest**
Presenter: **Jane-Ling Wang**, University of California Davis, United States
Co-authors: Hans-Georg Mueller

An introduction into the most commonly used methods of FDA. These include Functional Principal Component Analysis (FPCA) and the related concept of modes of variation, which is based on simple statistical notions such as mean and covariance function of a random process that can be inferred from the data. FPCA is an important dimension reduction tool and in sparse data situations can be used to impute functional data that are sparsely observed.Another core topic of FDA is functional regression, where one pairs functions or scalars as predictors with responses that are also functions or scalars. For the case where the predictors include functions, a difficult step that requires regularization is the inversion of a covariance operator, which is an ill-posed problem. Such an inverse problem is also related to some forms of functional correlation, which will be another core topic. Nonlinear methods have also found increasing interest. These include polynomial and quadratic regression relations, dimension reduction methods such as additive, continuously additive and index models, and other nonlinear approaches. Further topics of interest that may be covered are warping and manifold learning, the learning of time dynamics from observed realizations of the underlying stochastic process, multivariate and repeatedly observed functional data and stringing of high-dimensional data into functional data.

Saturday 27.08.2016 08:50 - 10:30 Parallel Session D – CRoNoS FDA 2016

TESTING IN MODELS WITH FUNCTIONAL DATA

CO011 **Room Sala 2** **Chair: Wenceslao Gonzalez-Manteiga**

CO0196: **Goodness-of-fit tests for the functional linear model based on randomly projected empirical processes.**
Presenter: **Manuel Febrero-Bande**, University of Santiago de Compostela, Spain
Co-authors: Juan A Cuesta-Albertos, Eduardo Garcia-Portugues, Wenceslao Gonzalez-Manteiga

Marked empirical processes, indexed by a randomly projected functional covariate, are considered to construct goodness-of-fit tests for the functional linear model with scalar response. The test statistics are built from continuous functionals over the projected process, resulting in efficient computational tests that exhibit root-n convergence rate and circumvent the curse of dimensionality. The weak convergence of the process is obtained conditionally on a random projection, whilst it is proved the almost surely equivalence between the testing for significance expressed on the original and on the projected functional covariate. The computation of the test in practise involves the calibration by wild bootstrap resampling and the combination of several p-values arising from different projections by means of the false discovery rate method. The finite sample properties of the test are illustrated in a simulation study for a variety of linear models, underlying processes and alternatives. R code that is available in the fda.usc library allows implementation and replication of the data applications provided.

CC0197: **Stationarity testing and break date estimation with functional time series**
Presenter: **Gregory Rice**, University of Waterloo, Canada

In fields ranging from economics and finance to energy research and climatology, data on certain continuous time phenomena are collected with high frequency. Often, such data can be parsed into natural, perhaps hourly or daily, segments that may be viewed as a time series of curves. A fundamental issue that must be addressed before an attempt is made to statistically model such data is whether these curves form a stationary functional time series. We will discuss the interpretation of stationarity in the context of function space valued random variables, and introduce testing procedures to test for stationarity with a given functional time series. The tests are developed as extensions of the broadly used tests in the KPSS family. When stationarity is rejected, it may be due to a number of factors, including a break or "change point" in the mean. We will also discuss some new methodology that does not rely on dimension reduction for estimating break dates with functional time series that contain a mean change.

CO0198: **Testing the influence of functional variables on functional responses**
Presenter: **Valentin Patilea**, CREST Ensai, France
Co-authors: Samuel Maistre

The considered problem is the test of the effect of Hilbert space-valued covariates on Hilbert space-valued responses. This general framework includes functional regression models checks against general alternatives, as well as testing conditional independence with functional data. The significance test for functional regressors in nonparametric regression with general covariates and scalar or functional responses is another example. We propose a new test based on kernel smoothing. The test statistic is asymptotically standard normal under the null hypothesis provided the smoothing parameter tends to zero at a suitable rate. The one-sided test is consistent against any fixed alternative and detects local alternatives a la Pitman approaching the null hypothesis. In particular we show that neither the dimension of the outcome nor the dimension of the functional covariates influences the theoretical power of the test against such local alternatives. Simulation experiments and a real data application illustrate the performance of the new test with finite samples.

CO0201: **Testing the predictor effect on a functional response**
Presenter: **Cesar Sanchez-Sellero**, University of Santiago de Compostela, Spain
Co-authors: Valentin Patilea, Matthieu Saumard

The problem of nonparametric testing for the no-effect of a random covariate (or predictor) on a functional response is considered. This means testing whether the conditional expectation of the response given the covariate is almost surely zero or not, without imposing any model relating response and covariate. The covariate could be univariate, multivariate or functional. Our test statistic is a quadratic form whose kernel is given by the inner product of the responses, weighted by univariate nearest neighbour smoothing weights. The asymptotic critical values are given by the standard normal law. When the covariate is multidimensional or functional, a preliminary dimension reduction device is used which allows the effect of the covariate to be summarized into a univariate random quantity. The test is able to detect not only linear but nonparametric alternatives. The responses could have conditional variance of unknown form and the law of the covariate does not need to be known. An empirical study with simulated and real data shows that the test performs well in applications.

SUMMER COURSE: SESSION III

CC038 **Room Sala 1** **Chair: Hans-Georg Mueller**

CK0206: **Functional data Analysis: From basics to current topics of interest**
Presenter: **Hans-Georg Mueller**, University of California Davis, United States
Co-authors: Jane-Ling Wang

An introduction into the most commonly used methods of FDA. These include Functional Principal Component Analysis (FPCA) and the related concept of modes of variation, which is based on simple statistical notions such as mean and covariance function of a random process that can be inferred from the data. FPCA is an important dimension reduction tool and in sparse data situations can be used to impute functional data that are sparsely observed.Another core topic of FDA is functional regression, where one pairs functions or scalars as predictors with responses that are also functions or scalars. For the case where the predictors include functions, a difficult step that requires regularization is the inversion of a covariance operator, which is an ill-posed problem. Such an inverse problem is also related to some forms of functional correlation, which will be another core topic. Nonlinear methods have also found increasing interest. These include polynomial and quadratic regression relations, dimension reduction methods such as additive, continuously additive and index models, and other nonlinear approaches. Further topics of interest that may be covered are warping and manifold learning, the learning of time dynamics from observed realizations of the underlying stochastic process, multivariate and repeatedly observed functional data and stringing of high-dimensional data into functional data.

Saturday 27.08.2016 12:00 - 13:15 Parallel Session F – CRoNoS FDA 2016

ROBUST FUNCTIONAL DATA ANALYSIS WITH APPLICATIONS

CO007 **Room Sala 2** **Chair: Ana Arribas-Gil**

CO0162: **Global and local functional depths**

Presenter: **Carlo Sguera**, Universidad Carlos III de Madrid, Spain

Co-authors: Pedro Galeano, Rosa Lillo

A functional data depth provides a center-outward ordering criterion which allows robust measures, such as the median, trimmed means, central regions or ranks, to be defined in the functional framework. A functional data depth can be global or local. With global depths, the degree of centrality of a curve x depends equally on the rest of the sample observations, while with local depths, the contribution of each observation in defining the degree of centrality of x decreases as the distance from x increases. We present a comparative analysis of the global and local approaches to the functional depth problem focusing on the "global" functional spatial depth (FSD) and its local version, the kernelized functional spatial depth (KFSD). First, we consider two illustrative real applications to show that FSD and KFSD may behave differently. Then, we present the results of a simulation study designed to understand when different behaviors between global and local depths should be expected.

CO0167: **An angle-based functional pseudo-depth for shape outlier detection**

Presenter: **Andre Rehage**, TU Dortmund University, Germany

Co-authors: Sonja Kuhnt

A measure especially designed for detecting shape outliers in functional data is presented. It is based on the tangential angles of the intersections of the centred data and its interpretation is the same as for a data depth. Due to its theoretical properties it is called functional tangential angle (FUNTA) pseudo-depth. Furthermore, a robustification called rFUNTA will be introduced. The existence of intersection angles is ensured through the centring. Assuming that shape outliers in functional data follow a different pattern than the regular data, the distributions of intersection angles differ from each other. Robustness properties of the two introduced measures and their performances in real data sets as well as simulation studies are investigated.

CO0155: **Finding outliers in surface data and video**

Presenter: **Peter Rousseeuw**, KU Leuven, Belgium

Co-authors: Mia Hubert, Jakob Raymaekers, Pieter Segaert

Surface and image data can be considered as functional data with a bivariate domain. It is well known that classical statistical techniques to detect outlying surfaces or to flag outlying parts of a surface may themselves be distorted by the outliers. Outlyingness in multivariate data can be measured better by a projection pursuit approach. A new display is proposed, based on the mean and the variability of the degree of outlyingness at each grid point. A rule is constructed to flag the outlying functions in the resulting functional outlier map. Heatmaps of their outlyingness indicate the regions which are most deviating from the regular surfaces. The method works on univariate functional data as well as multivariate functional data. To illustrate the method it is applied to fluorescence excitation-emission spectra after fitting a PARAFAC model, to MRI image data which are augmented with their gradients, and to video surveillance data.

METHODS FOR FUNCTIONAL RESPONSE MODELS

CO025 **Room Sala 1** **Chair: Jeff Goldsmith**

CO0182: **Boosting flexible functional regression models with a high number of functional historical effects**

Presenter: **Sarah Brockhaus**, Ludwig-Maximilians-University Munich, Germany

Co-authors: Michael Melcher, Friedrich Leisch, Sonja Greven

A flexible framework for function-on-function regression models with historical effects is proposed. For functional response and covariate both observed over the same time interval, a historical functional effect induces an association between response and covariate such that only past values of the covariate influence the current value of the response. Effects with more general integration limits can be specified including for example lag and lead effects. Further covariate effects like linear and smooth effects of scalar covariates can be included in the models. Estimation is conducted by a component-wise gradient boosting algorithm which allows to model different features of the response distribution, e.g. the expectation or a quantile, by minimizing the according loss function. Moreover, boosting can estimate models in high dimensional data settings, with more covariates than observations, and inherently does variable selection. The R add-on package FDboost provides an open-source implementation of the methods. The motivating application is a biotechnological dataset of 25 Escherichia coli fermentations for the production of a model protein. For monitoring of the process, the cell dry mass, which is time-consuming to determine, should be predicted during new fermentations using past and current values of a potentially large number of easily accessible process variables.

CO0181: **Modeling heterogeneity in motor learning using heteroskedastic functional principal components**

Presenter: **Daniel Backenroth**, Columbia Mailman School of Public Health, United States

Co-authors: Jeff Goldsmith

A novel method is proposed for estimating population-level and subject-specific effects of covariates on the variability of functional data. We extend the functional principal components analysis framework by modeling the variance of principal component scores as a function of covariates and subject-specific random effects. In a setting where principal components are largely invariant across subjects and covariate values, modeling the variance of these scores provides a flexible and interpretable way to explore factors that affect the variability of functional data. The motivation arises from a novel dataset from an experiment assessing upper extremity motor control, and quantifies the reduction in motion variance associated with skill learning.

CO0183: **Functional regression for investigating the feeding behavior of animals**

Presenter: **Jan Gertheiss**, Clausthal University of Technology, Germany

Co-authors: Sonja Greven, Engel Hessel, Verena Maier, Fabian Scheipl

A group of pigs is observed over a period of about 100 days. Using high frequency radio frequency identification, it is recorded when each pig is feeding, leading to very dense sequences of binary observations for each pig and day. Goals of the data analysis are to find pig-specific feeding profiles showing the typical feeding pattern of each pig, and to make short-term predictions of pig-specific feeding probabilities. Different approaches for modeling the data are discussed: (1) a marginal functional logistic regression approach modeling the binary measurements by assuming latent, smooth and cyclic pig-specific profiles. To account for correlation of measurements, robust standard errors and corresponding pointwise confidence intervals can be used. As an alternative, (2) a conditional model including pig-specific functional random effects or lagged responses to account for within-pig correlation is considered. By contrast to the marginal model, the latter model also allows for short-term predictions of feeding behavior.

Saturday 27.08.2016 15:00 - 16:40 Parallel Session G – CRoNoS FDA 2016

RECENT ADVANCES IN FUNCTIONAL DATA ANALYSIS

CO015 **Room Sala 2** **Chair: Alois Kneip**

CO0178: **Nonparametric registration to low-dimensional function spaces**
Presenter: **Alois Kneip**, University of Bonn, Germany
Co-authors: Heiko Wagner

Registration aims to decompose amplitude and phase variation of samples of curves. Phase variation is captured by warping functions which monotonically transform the domains. Resulting registered curves should then only exhibit amplitude variation. Most existing registration method rely on aligning typical shape features like peaks or valleys to be found in each sample function. It is shown that this is not necessarily an optimal strategy for subsequent statistical data exploration and inference. In this context a major goal is to identify low dimensional linear subspaces of functions that areable to provide accurate approximations of the observed functional data. We present a registration method where warping functions are defined in such a way that the resulting registered curves span a low dimensional linear function space. Problems of identifiability are discussed in detail, and connections to established registration procedures are analyzed. Together with a suitable analysis of warping functions, the method allows to decompose functional data in a way that might be more informative than standard functional principal component analysis. We derive inference results for situations, where the true functions have to be reconstructed from discrete, noisy observations. The method is applied to real and simulated data.

CO0177: **Reconstruction of partially observed (non-)sparse functional data**
Presenter: **Dominik Liebl**, University Bonn, Germany
Co-authors: Alois Kneip

A new prediction procedure is proposed that allows to reconstruct functional data from their fragmental observations. Additionally, we derive verifiable conditions under which our prediction procedure allows for a perfect reconstruction without any prediction error. Similarly to the context of sparse functional data, it is assumed that only noisy discretization points of the random functions are observable. By means of a double asymptotic we consider the asymptotic mean square prediction error of our functional PCA based estimator for all cases from sparsely to densely sampled discretization points per function. Applicability of our prediction method is demonstrated by a real data study in which we seek to reconstruct electricity price functions.

CO0168: **Regression imputation in the functional linear model with a response missing at random**
Presenter: **Christophe Crambes**, University of Montpellier, France
Co-authors: Yousri Henchiri

Functional linear regression is considered when some observations of the real response are missing, while the functional covariate is completely observed. A regression imputation method of missing data is presented, using functional principal component regression to estimate the functional coefficient of the model. We study the asymptotic behaviour of the error we commit when the missing data is replaced by the regression imputed value, in a 'missing at random' framework. The practical behaviour of the method is also studied on simulated data sets, and a real case study for which we forecast a pollution indicator using a temperature curve in temperate cities of France.

CO0172: **On the CLT for discrete Fourier transforms of functional time series**
Presenter: **Siegfried Hoermann**, Univ libre de Bruxelles, Belgium
Co-authors: Clement Cerovecki

A strictly stationary functional time series is considered and the weak convergence of its discrete Fourier transforms is studied under sharp conditions. As a side result we obtain the regular CLT for partial sums under mild assumptions.

SUMMER COURSE: SESSION IV

CC039 **Room Sala 1** **Chair: Hans-Georg Mueller**

CK0207: **Functional data Analysis: From basics to current topics of interest**
Presenter: **Hans-Georg Mueller**, University of California Davis, United States
Co-authors: Jane-Ling Wang

An introduction into the most commonly used methods of FDA. These include Functional Principal Component Analysis (FPCA) and the related concept of modes of variation, which is based on simple statistical notions such as mean and covariance function of a random process that can be inferred from the data. FPCA is an important dimension reduction tool and in sparse data situations can be used to impute functional data that are sparsely observed.Another core topic of FDA is functional regression, where one pairs functions or scalars as predictors with responses that are also functions or scalars. For the case where the predictors include functions, a difficult step that requires regularization is the inversion of a covariance operator, which is an ill-posed problem. Such an inverse problem is also related to some forms of functional correlation, which will be another core topic. Nonlinear methods have also found increasing interest. These include polynomial and quadratic regression relations, dimension reduction methods such as additive, continuously additive and index models, and other nonlinear approaches. Further topics of interest that may be covered are warping and manifold learning, the learning of time dynamics from observed realizations of the underlying stochastic process, multivariate and repeatedly observed functional data and stringing of high-dimensional data into functional data.

Saturday 27.08.2016 17:10 - 18:50 Parallel Session H – CRoNoS FDA 2016

METHODOLOGICAL AND APPLIED CONTRIBUTIONS ON FUNCTIONAL DATA ANALYSIS

CO005 **Room Sala 2** **Chair: M Carmen Aguilera-Morillo**

CO0156: **Model-based co-clustering for functional data**

Presenter: **Julien Jacques**, University Lyon II, France

Co-authors: Yosra Ben Slimen, Sylvain Allio

Nowadays, mobile telecommunication is growing rapidly which creates new challenges for mobile operators. Despite this fast evolution, many applications linked to network management are still achieved manually, such as network troubleshooting and optimization. To accomplish theirs tasks, technicians refer basically to key performance indicators (KPIs), distinguished by their functional behaviors. We aim to analyze those KPIs, in order to help the technical support team with their daily decision making and to add additional knowledge for self-organizing networks. Our dataset is composed of days as observations and KPIs as attributes. Therefore, each cell ij of the dataset is a curve of the evolution of the KPI j during the day i. Our goal is to extract the relationships between KPIs and days with a co-clustering strategy, and thus to discover the double structure hidden in the dataset. For this, a model-based co-clustering algorithm for functional data is proposed. We use the latent block model with a Gaussian assumption for modeling the functional principal component scores of the curves. Parameter inference is carried out thanks to a SEM-Gibbs algorithm. Our model is generic; it can be applied on different use cases. Moreover, it can be used as a preprocessing step for network management. It can also be a tool to create supervised datasets that will help in network troubleshooting and optimization tasks.

CO0173: **Forecasting functional time series in electricity markets with a seasonal ARIMAX Hilbertian model**

Presenter: **Jose Portela**, Universidad Pontificia Comillas, Spain

Co-authors: Antonio Munoz, Estrella Alonso

A functional time series is the realization of a stochastic process where each observation is a continuous function defined in a finite interval. In order to forecast these functions, a seasonal ARIMAX Hilbertian model is presented. It extends the structure of ARIMA models to functional data using integral operators in the L2 space. The kernels of operators are modeled as linear combinations of sigmoid functions, where the parameters of each sigmoid are estimated using a Quasi-Newton algorithm which minimizes the sum of squared errors. This functional model allows forecasting functional time series taking into account time dependencies (autoregressive and moving average terms), seasonality as well as scalar and functional exogenous variables. An empirical study is presented for the time series of hourly residual demand curves in the Spanish day-ahead electricity market. The residual demand curves model the competitive behavior of the agents bidding in an electricity market. For every auction, the residual demand is defined as the clearing price of the market expressed as a function of the amount of energy an agent is able to buy or sell. Being able to forecast these curves is the first and essential step in the design of optimal bidding strategies.

CO0166: **The application of functional linear models on remote sensing data in oceanography**

Presenter: **Nihan Acar-Denizli**, Mimar Sinan Fine Arts University, Turkey

Co-authors: Gulay Basarir, Isabel Caballero de Frutos, Pedro Delicado

With the development of technology, functional data analysis gain importance in analysing large data sets where the observations are sampled over a dense time interval, space or on a spectrum that consists of different frequency channels. Remote Sensing (RS) data obtained from satellites are an example of spectral data. Particulary in oceanography, RS data are used to predict ocean characteristic parameters such as Sea Surface Temperature (SST), Chlorophyll-a content (Chl-a) and Total Suspended Solids (TSS). Different functional linear regression models for scalar responses are applied on the Remote Sensing (RS) data obtained from full spatial resolution (FRS) MEdium Resolution Imaging Spectrometer (MERIS) on board the Envisat multispectral platform. The purpose is to predict the amount of TSS in the coastal zone adjacent to the Guadalquivir estuary. The models are compared by using Mean Error of Prediction (MEP) computed from Leave One Out Cross Validation (LOOCV).

CO0185: **Variable selection in the functional linear concurrent model**

Presenter: **Jeff Goldsmith**, Columbia University, United States

Co-authors: Joseph Schwartz

Methods for variable selection are proposed in the context of modeling the association between a functional response and concurrently observed functional predictors. This data structure, and the need for such methods, is exemplified by our motivating example: a study in which blood pressure values are observed throughout the day, together with measurements of physical activity, location, posture, attitude, and other quantities that may influence blood pressure. We estimate the coefficients of the concurrent functional linear model using variational Bayes and jointly model residual correlation using functional principal components analysis. Latent binary indicators partition coefficient functions into included and excluded sets, incorporating variable selection into the estimation framework. The proposed methods are evaluated in real-data analyses, and are implemented in a publicly available R package.

SUMMER COURSE: SESSION V

CC040 **Room Sala 1** **Chair: Hans-Georg Mueller**

CK0205: **Functional data Analysis: From basics to current topics of interest**

Presenter: **Hans-Georg Mueller**, University of California Davis, United States

Co-authors: Jane-Ling Wang

An introduction into the most commonly used methods of FDA. These include Functional Principal Component Analysis (FPCA) and the related concept of modes of variation, which is based on simple statistical notions such as mean and covariance function of a random process that can be inferred from the data. FPCA is an important dimension reduction tool and in sparse data situations can be used to impute functional data that are sparsely observed.Another core topic of FDA is functional regression, where one pairs functions or scalars as predictors with responses that are also functions or scalars. For the case where the predictors include functions, a difficult step that requires regularization is the inversion of a covariance operator, which is an ill-posed problem. Such an inverse problem is also related to some forms of functional correlation, which will be another core topic. Nonlinear methods have also found increasing interest. These include polynomial and quadratic regression relations, dimension reduction methods such as additive, continuously additive and index models, and other nonlinear approaches. Further topics of interest that may be covered are warping and manifold learning, the learning of time dynamics from observed realizations of the underlying stochastic process, multivariate and repeatedly observed functional data and stringing of high-dimensional data into functional data.

Sunday 28.08.2016 08:50 - 10:50 Parallel Session J – CRoNoS FDA 2016

CONTRIBUTIONS ON METHODOLOGICAL AND APPLIED FUNCTIONAL DATA ANALYSIS

CG006 **Room Sala 2** **Chair: Ana M Aguilera**

CC0158: **Dependent generalized functional linear models**

Presenter: **Sneha Jadhav**, Michigan State University, United States

A framework is proposed for regression models where the dependent variable is a scalar with certain dependence structure and the independent variable is a function. In particular we assume that the data consists of clusters that have dependence within each cluster but are independent with respect to each other. We use generalized estimating equations to estimate the underlying parameters and establish their joint asymptotic normality. This asymptotic distribution is used to test asymptotically the significance of the dependent variable on the response variable. We apply these results on a family gene sequencing data. Individuals between families are independent but may be dependent within a family, thus necessitating for a method with above properties. Our simulations indicate that under certain conditions, functional approach has higher power for the high dimensional sequencing data as compared some current popular approaches.

CC0160: **Prediction of functional moving average models**

Presenter: **Johannes Klepsch**, Technical University Munich, Germany

Co-authors: Claudia Klueppelberg

First, a fully functional representation is given for the linear one-step predictor of the functional moving average (FMA) model in terms of the past of the process. Assuming invertibility of the process, we derive asymptotic properties of the operators involved in the representations of the predictors. Then, as the infinite dimensionality of the model prevents us from applying for example the innovation algorithm to compute the predictors, we project the FMA model on an arbitrary K-dimensional subspace. In contrast to the functional autoregressive model, we show that the projected FMA model still follows the dynamics of a FMA model of the same order or less, with a new K-dimensional innovation process. We show that we get arbitrarily close to the original FMA model by increasing K and give implications for the prediction of FMA models.

CC0176: **A functional regression approach for modelling the retinal nerve fiber layer thickness in the eyes of healthy subjects**

Presenter: **Eleonore Pablik**, Medical University of Vienna, Austria

Co-authors: Florian Frommlet

One hundred healthy volunteers underwent ophthalmic examination where the retinal nerve fiber layer (RNFL) thickness was measured in a circle of 3.4 mm diameter around the optic disc with Fourier-domain optical coherence tomography. Measurements were obtained in 256 equidistant sectors on the circle. Functional data analysis was used to develop a model to partly explain the inter-subject variability of RNFL in these healthy subjects in order to obtain a narrower range of normative RNFL data. The long term goal of this modelling is to improve the diagnosis of early glaucoma or other diseases affecting the RNFL thickness. As a first step we used landmarking to provide adjusted percentiles for each of the 256 points of measurement instead of the raw data's percentiles. In the second step we modelled the patient-specific variation around the adjusted median with 16 basis functions. Due to the fact that a positive correlation between retinal nerve layer thickness and the amount of retinal vessels could already be shown in earlier publications we tried to explain the magnitude of each of these basis functions with the amount of retinal vessels in the corresponding area using linear regression models. We compare our functional regression approach with a standard multiple regression approach previously used in the literature.

CC0186: **Detection of periodicity in functional time series**

Presenter: **Gilles Nisol**, ULB, Belgium

Determining whether the dynamics of certain phenomena has a periodic pattern is essential in many fields, notably in climatology, finance or environmental sciences. For example, a weekly pattern in pollution level curves may be associated with an anthropogenic influence. The aim is to extend the range of application of existing periodicity tests to time series of functional data. We will investigate two approaches for testing whether observations are stationary, against the alternative hypothesis that there is a fixed periodic trend inherent. First, we will consider the projection of the time series onto a finite subspace and derive the likelihood ratio test statistic for the projected series. Second, we will define a purely functional test, in the sense that it does not require any projection on a subspace. To start with, our theory is derived in an i.i.d. Gaussian framework with a simple sinusoidal alternative, and then it is generalized to stationary observations and some arbitrary periodicity pattern. We will then conduct a simulation study in order to assess the power functions of our tests. Finally, we apply our procedure to intraday volatility curves of SP100 which are shown to exhibit a weekly periodicity.

CC0195: **Distribution-free interval-wise inference for functional-on-scalar linear models**

Presenter: **Sara Sjostedt de Luna**, Umea University, Sweden

Co-authors: Konrad Abramowicz, Charlotte Hager, Alessia Pini, Lina Schelin, Simone Vantini

A distribution-free procedure is introduced for testing a functional-on-scalar linear model with fixed effects. The procedure does not only test the global hypothesis on all the domain, but also selects the intervals where statistically significant effects are detected. We prove that the proposed tests are provided with an asymptotic control of the interval-wise error rate, i.e. the probability of falsely rejecting any interval of true null hypotheses. The procedure is then applied to one-leg hop data from a study on anterior cruciate ligament injury. We compare knee kinematics of three groups of individuals (two injured groups with different treatments, and one group of healthy controls), taking individual-specific covariates into account.

SUMMER COURSE: SESSION VI

CC041 **Room Sala 1** **Chair: Hannu Oja**

CK0209: **Independent component analysis for functional data**

Presenter: **Hannu Oja**, University of Turku, Finland

In independent component analysis (ICA) it is assumed that the observed random vectors are linear combinations of latent, mutually independent random variables called the independent components. It is then often assumed that only the non-Gaussian independent components are of interest and the Gaussian components are treated as noise. The aim is to extract, using an affine transformation, the non-Gaussian components as well as to isolate, using a projection, the signal and noise subspaces. In this way, ICA can recover the patterns that cannot be identified by classical principal component analysis (PCA). Extensions to functional data where observed units are random functions rather than random vectors are discussed. The independent component model assumes, e.g., that the random function X can be decomposed as $X = Z + E$ where E is Gaussian and Z is finite-dimensional such that, for some d and some functions $f_1, ..., f_d$, $Z \sim \sum_{i=1}^{k} Z_i f_i$ with independent real valued random variables $Z_1, ..., Z_d$. One then has first (i) to find a decomposition $X = Z + E$, that is, candidates for signal space and noise space, and then (ii) find and separate the Gaussian and non-Gaussian independent components $Z_1, ..., Z_d$ as well as functions $f_1, ..., f_d$. The extension of FOBI is considered in some details and illustrated with simulated and real datasets.

Sunday 28.08.2016 11:20 - 13:00 Parallel Session K – CRoNoS FDA 2016

STATISTICS IN HILBERT SPACES

CO013 **Room Sala 1** **Chair: Gil Gonzalez-Rodriguez**

CO0163: **Sparse regularization for functional logistic models and its application to time-dependent biomarker detection**
Presenter: **Hidetoshi Matsui**, Shiga University, Japan
Co-authors: Mitsunori Kayano, Seiya Imoto, Rui Yamaguchi, Satoru Miyano
We consider constructing functional logistic regression models based on sparse regularization in order to identify genes with dynamic alterations in case/control study. We employ the mixed effects model based on basis expansions in order to convert time course profiles measured at small and irregularly spaced time points into functional data, and then estimate the functional logistic regression model using the penalized likelihood method with the group elastic net penalty in order to select genes that have time course expression profiles. We compare the efficacy of the proposed method to that of the existing method through Monte Carlo simulations, and then show through gene expression data that our method detects dynamic alterations of genes, which cannot be found by the existing method.

CO0164: **Data depth for measurable random mappings**
Presenter: **Stanislav Nagy**, KU Leuven, Belgium
Data depth is a mapping, which to a point in a multivariate vector space $s \in S$ and a probability measure P on S assigns a number $D(s;P)$ describing how central s is with respect to P, in an attempt to generalize quantiles to multivariate data. For S having infinite dimension, depth is typically considered only for S being the Banach space of continuous functions over a compact interval. We explore possibilities of extension of known depth functionals beyond this simplest setting. We discuss definitions, theoretical properties, and consistency/measurability issues connected with a straightforward generalization of commonly used depth functionals towards multidimensional random mappings which may lack continuity, be observed discretely, or be contaminated with additive noise.

CO0169: **Sparse high-dimensional and functional autoregressions**
Presenter: **Xinghao Qiao**, London School of Economics, United Kingdom
Co-authors: Shaojun Guo
Multivariate functional data arise in a broad spectrum of real applications. However, many studies in high dimensional functional data focus primarily on the critical assumption of independent and identically distributed (i.i.d.) samples. We consider a sparse high dimensional functional autoregressive model to characterize the dynamic dependence across different functional time series. We propose a new regularization method via group lasso to estimate the autoregressive coefficient functions and derive non-asymptotic bounds for the estimation errors of the regularized estimates. We also introduce a measure for stationary functional processes that provides insight into the effect of dependence on the accuracy of the regularized estimates. Finally, we show that the proposed model significantly outperforms its competitors through both simulated and real data sets.

CO0193: **Location M-estimation approach for functional data**
Presenter: **Beatriz Sinova**, University of Oviedo, Spain
Co-authors: Gil Gonzalez-Rodriguez, Stefan Van Aelst
M-estimators are a well-known and successful methodology to summarize the centre of univariate or multivariate real-valued data. The M-estimation approach is extended to the functional-valued setting. First, some developments from the literature for robust nonparametric density estimation are applied in order to define location M-estimators for functional data and establish conditions for their existence and uniqueness. It is shown that well-known loss functions such as the Huber, Hampel and Tukey ones yield robust location estimators also in the functional context. Then, the strong consistency and robustness, by means of both the finite sample breakdown point and the influence function, of the estimators of this proposal are analyzed. Finally, some simulations show the finite-sample performance of functional location M-estimators with respect to other robust approaches like trimmed means.

FUNCTIONAL DATA MODELLING

CO017 **Room Sala 2** **Chair: Jean-Baptiste Aubin**

CO0170: **Generalized functional linear models under choice-based sampling**
Presenter: **Sophie Dabo**, University-Lille, France
Co-authors: Mohamed Salem Ahmed
A functional binary model in a context of sampling data is proposed and estimated. This problem is known respectively in econometric and epidemiology literatures, as Choice-Based Sampling and case-control study, in discrete choice model. Unlike the random sample where all items in the population have the same probability of being chosen, the Choice-Based Sampling (CBS) in discrete choice model is a type of sampling where the classification of the population into subsets to be sampled is based on the choices or outcomes. In practice, it could be of interest to model choice of individuals using some functional covariates instead of real valued random variables. To this end, we introduce the Choice-Based sampling in a functional framework (functional generalized linear models). We reduce the infinite dimensional of the space of the explanatory random function using a Karhunen-Loeve expansion and maximize a conditional likelihood function. Our method is based on the components of a Functional Principal Components Analysis adapted to the context of Choice-Based Sampling. Asymptotic properties of our estimate are given. We present some simulated experiments including genetic data, to investigate the finite sample performance of the estimation method. The potential of the functional choice-based sampling model to integrate the special non-random features of the sample, that would have been hard to see otherwise is also outlined.

CO0190: **Quantile regression for functional data**
Presenter: **Maria Franco Villoria**, University of Torino, Italy
Co-authors: Rosaria Ignaccolo, Marian Scott
Quantile regression allows estimation of the relationship between response and explanatory variables at any percentile of the distribution of the response (conditioned on the explanatory variables). The idea is to extend quantile regression to the functional case, where the observational unit is no longer a point but a whole curve. We rewrite the quantile regression model as a generalized additive model where both the functional covariates and the functional coefficients are parametrized in terms of B-splines. Parameter estimation would involve minimizing a sum of asymmetrically weighted absolute deviations, for which linear programming methods are used. Instead, a weighted least squares approach is preferred; by approximating the absolute residuals with the squared residuals (and adjusting the weights accordingly), model parameters can be estimated using a penalized iterative reweighed least squares (PIRLS) algorithm. We evaluate the performance of the model by means of a simulation study, in which different levels of noise and different functional forms for the corresponding functional coefficients are considered.

CO0191: **Functional principal components for concentration curves**
Presenter: **Enea Bongiorno**, Universita del Piemonte Orientale, Italy
Co-authors: Aldo Goia

Concentration 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 that refer to the objects oriented data analysis literature. In fact, the family of concentration curves lacks of very basilar structures like the vectorial one and, hence, should be treated with ad hoc methods. The aim is to take care of such lacks providing a rigorous functional framework for concentration curves where it is possible to define Functional Principal Component Analysis (FPCA). The latter technique is then implemented and used to explore functional dataset.

CC0199: **Nonparametric time-simultaneous inference for transition probabilities in multi-state Markov models**

Presenter: **Dennis Dobler**, University of Ulm, Germany

Co-authors: Markus Pauly, Tobias Bluhmki, Jan Beyersmann

The analysis of transition probability matrices of non-homogeneous Markov processes is of great importance (especially in medical applications) and it constantly gives rise to new statistical developments. For each individual, the collected functional data consists of a state function indicating its state occupations in time. While observations may be incomplete, e.g. due to random left-truncation and right-censoring, estimation of these probabilities is conducted by employing counting processes and the Aalen-Johansen estimator. However, results of weak convergence towards a Gaussian process cannot be utilized straightforwardly for time-simultaneous inference due to complicated limiting covariance structures revealing a lack of tabulated quantiles. In order to construct asymptotically valid confidence bands, we introduce a flexible class of resampling techniques using a multiplier bootstrap. For general multi-state Markov models the wild bootstrap is shown to be consistent due to martingale arguments. As an intermediate result we obtain the wild bootstrap consistency for the Nelson-Aalen estimator of cumulative hazard functions. In both cases conditional weak convergence towards Gaussian processes with correct covariance functions result in consistent tests and confidence bands. Simulations indicate that Poi(1)-1 wild bootstrap multipliers may improve the type I error control in comparison to standard normal multipliers, which are typically used in similar inference problems.

Sunday 28.08.2016 14:30 - 16:10 Parallel Session L – CRoNoS FDA 2016

NON- AND SEMI-PARAMETRIC APPROACHES IN FUNCTIONAL STATISTICS

CO009 **Room Sala 1** **Chair: Enea Bongiorno**

CO0161: **Modeling space-time functional data with complex dependencies via regression with partial differential regularizations**
Presenter: **Mara Sabina Bernardi**, Politecnico di Milano, Italy
Co-authors: Laura Sangalli, Gabriele Mazza, James Ramsay

A semiparametric model for the analysis of functional data with complex dependencies is presented. The data considered can be seen as spatially dependent curves or time dependent surfaces. The model is based on the idea of regression with partial differential regularizations. Among the various modeling features, the proposed method is able to deal with spatial domains featuring peninsulas, islands and other complex geometries. Space-varying covariate information is included in the model via a semiparametric framework. The estimators have a penalized regression form, they are linear in the observed data values, and have good inferential properties. The use of numerical analysis techniques, and specifically of finite elements, makes the model computationally very efficient. The model is compared via simulations to other spatio-temporal techniques and it is illustrated via an application to the study of the annual production of waste in the municipalities of Venice province.

CO0184: **On a regression model with constraints in Hilbert spaces**
Presenter: **Marta Garcia Barzana**, University of Oviedo, Spain
Co-authors: Ana Colubi, Gil Gonzalez-Rodriguez

The least-squares estimation of linear regression models involves an optimization problem that may be subject to a certain group of constraints. The well-known constrained least-squares approach assumes that the number of inequality linear constrains is fixed. This framework will be extended by removing such an assumption. Thus, the number of constrains can vary depending on the sample size. This problem has been addressed in the context of linear regression with interval data. However, the goal is to extend the problem to the abstract case of regression models in Hilbert spaces, which accommodates as well more complex data, such as functional data. An estimator is proposed and a case-based example is presented.

CC0192: **Testing optimal dimension for suitable Hilbert-valued processes**
Presenter: **Jean-Baptiste Aubin**, Insa-Lyon, France
Co-authors: Enea Bongiorno

The small-ball probability (SmBP) of a Hilbert-valued process is considered. Recent works have shown that, for a fixed number d and as the radius epsilon of the ball tends to zero, the SmBP is asymptotically proportional to (a) the joint density of the first d principal components (PCs) evaluated at the center of the ball, (b) the volume of the d-dimensional ball with radius epsilon, and (c) a correction factor weighting the use of a truncated version of the process expansion. Under suitable assumptions on the decay rate of the eigenvalues of the covariance operator of the process, it has been shown that the correction factor in (c) tends to 1 as the dimension increases. The properties of the correction factor are studied and a consistent estimator is introduced. Features of such estimator allow to conservatively test whenever the correction factor equals 1. This implicitly implies that, for the class of processes whose eigenvalues of the covariance operator decay hyper-exponentially, an optimal dimension can be defined allowing to use a "finite-dimensional" approach in approximating the SmBP and, hence, providing a natural model advantage.

CC0188: **Estimation of functional sparsity in varying coefficient models with functional covariates**
Presenter: **Juhyun Park**, Lancaster University, United Kingdom
Co-authors: Catherine Tu, Juhyun Park, Haonan Wang

Nonparametric estimation is studied under a certain type of sparsity consideration for varying coefficient models in analyzing longitudinal data with functional covariates. The problem of sparse estimation is well understood in the parametric setting as variable selection. Sparsity is the recurrent theme that also encapsulates interpretability in the face of high dimensional regression problems. For nonparametric models, interpretability not only concerns the number of covariates involved but also the functional form of the estimates, so the sparsity consideration is much more complex. To distinguish the types of sparsity in nonparametric models, we call the former "global sparsity" and the latter "local sparsity", which constitute "functional sparsity". Most existing methods focus on directly extending the framework of parametric sparsity for linear models to nonparametric function estimation to address one or the other, but not both. We develop a penalized estimation procedure that simultaneously addresses both types of sparsity in a unified framework. We establish asymptotic properties of estimation consistency and sparsistency of the proposed method. Our method is illustrated in simulation study and real data analysis, and is shown to outperform the existing methods in identifying both local sparsity and global sparsity.

FUNCTIONAL DATA ANALYSIS: APPLICATIONS

CC003 **Room Sala 2** **Chair: Gregory Rice**

CC0159: **Two R-packages for object-oriented functional data analysis**
Presenter: **Clara Happ**, LMU Munich, Germany

Based on the philosophy of object-oriented data analysis, the presented `R`-package `funData` implements functional data in a natural and intuitive manner. The data is modeled as an S4 class, using simply the argument values (e.g. time) and the functional observations. No additional assumptions, e.g. on a basis function representation, have to be made. Supported data types are univariate functional data on one- or higher dimensional domains, data on irregular grids and multivariate functional data. The core methods of the package include visualization and arithmetics (function/function and function/scalar) for all data types. Further, it provides a full simulation toolbox for univariate and multivariate functional data on one- and two-dimensional domains. The `MFPCA` package allows to calculate principal component analysis for multivariate functional data on different dimensional domains. It is built on the `funData` package and gives easily interpretable and theoretically founded results based on a Karhunen-Loeve representation. Both packages are available online, including examples and an extensive documentation.

CC0194: **Multi-resolution clustering of time dependent functional data with applications to climate reconstruction**
Presenter: **Konrad Abramowicz**, Umea University, Sweden
Co-authors: Lina Schelin, Sara Sjostedt de Luna, Johan Strandberg

A multi-resolution approach is presented to clustering dependent functional data. Given a lattice of (time) points, a function is observed at each grid point. We assume that there are latent (unobservable) groups that vary with over time. We consider the case when at different time scales (resolutions) different groupings arise, with groups being characterised by distinct frequencies of the observed functions. We propose and discuss a non-parametric double clustering based method, which identifies latent groups at different scales. We present an application of the introduced methodology to varved lake sediment, data aiming at reconstructing winter climatic regimes in northern Sweden at different resolutions during the last six thousand years.

CC0189: **Functional data methods for clustering pigs according to their feeding behavior**
Presenter: **Annette Moeller**, University of Goettingen, Germany
Co-authors: Jan Gertheiss, Engel Hessel

Subject-specific, high resolution registrations of pigs are considered at the trough across one fatting period of about 100 days. By means of high

frequency radio frequency identification (HF RFID) it was recorded when each pig is feeding. In a previous work, a marginal functional logistic regression model has been used to estimate smooth, pig-specific feeding profiles from the binary functional data available. These profiles reveal the typical feeding pattern of each pig over the day. Our objective is now to explore structures in the feeding profiles, identifying groups among the pigs with similar feeding behavior by using suitable methods that account for the functional nature of the data. In the case study presented we explore and compare popular clustering algorithms and especially the use of different (functional) distance or dissimilarity measures. Preliminary results indicate that the pigs can show quite different feeding behavior, defining some distinct groups. While the 'natural' two-peak profile is prominent, there are also a lot of profiles departing from this typical behavior, as for example one-peak profiles or profiles showing no distinct peak. Having identified a reasonable cluster structure, we proceed by investigating whether the group membership has a significant influence on the weight (gain) of the respective pigs.

CP0180: **Density functions from the exponential family as units of Bayes space: A simulation study**

Presenter: **Renata Talska**, Palacky University Olomouc, Czech Republic

Co-authors: Karel Hron, Alessandra Menafoglio

Probability density functions (PDFs) form a special class of functional data, carrying primarily relative information and characterized by specific features like scale invariance and relative scale. The Bayes spaces have been recently developed to capture the relative nature of PDFs by embedding the statistical analysis into an appropriate separable Hilbert space. Density functions from the exponential family form affine subspaces of a Bayes space, whose dimension honours the number of parameters of the distribution. The aim is to analyze possibility of representing and reducing the dimensionality of densities from the exponential family using the Bayes space methodology. For instance, we will show that proper choice of parameters is needed to single out the correct dimensionality of the considered family. Furthermore, functional principal component analysis in the Bayes spaces is employed to reveal interesting features of PDFs from the exponential family. A mapping of PDFs from Bayes spaces to the L2 space (i.e. centred logratio transformation) enables one to work according to the Bayes space geometry, by applying familiar tools of functional data analysis (e.g. FPCA).

Authors Index

23rd International Conference on Computational Statistics
COMPSTAT 2018

Iasi, Romania, August 28-31, 2018