PROGRAMME AND ABSTRACTS

CRoNoS Workshop and Spring Course on

Multivariate Data Analysis and Software (CRoNoS MDA 2018)

http://cmstatistics.org/CRONOSMDA2018/

Poseidonia Beach Hotel, Limassol, Cyprus
3-5 April 2018

Dear Friends and Colleagues,

We welcome you warmly to Limassol, for the CRoNoS Workshop and Spring Course on *Multivariate Data Analysis and Software* (CRoNoS MDA 2018). These events are co-organized by the COST Action *Computationally-intensive and RObust analysis of NOn-Standard data* (CRoNoS), the Cyprus University of Technology and the Frederick University. CRoNoS is a network of over 80 European researchers spanning computing, statistics, machine learning, and mathematics. Their aim is to develop new models, methods and efficient, numerically stable, and well-conditioned robust strategies to improve knowledge extraction from non-perfect and non-standard datasets. More information is available at http://www.cronosaction.com/.

The CRoNoS MDA 2018 programme consists of a course of 18 hours complemented with 4 plenary talks, 9 session and about 25 presentations. The CRoNos Chairs have endeavoured to provide a balanced and stimulating programme that will appeal to the diverse interests of the participants. It is hoped that the conference venue will provide the appropriate environment to enhance your contacts and to establish new ones. We acknowledge the support of our hosts and sponsors, and particularly the Cyprus University of Technology, the Frederick University and the COST office.

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

Looking forward, the closing events of the CRoNoS Action will be held again in Limassol, from Sunday 14th to Tuesday 16th of April 2019. You are invited and encouraged to actively participate in these events.

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

The CRoNos Chairs
Ana Colubi and Erricos J. Kontoghiorghes.

SOCIAL EVENTS

- *The coffee breaks* will take place at the Foyer of the Mezzanine of the Poseidonia Beach Hotel. You must have your conference badge in order to attend the coffee breaks.
- *Welcome Reception, Tuesday 3rd of April 2018, from 20:00-21:30.* The Welcome Reception is open to all registrants and accompanying persons who have purchased a reception ticket. It will take place at the Poseidon terrace of the Poseidonia Beach Hotel. Participants must bring their conference badge.
- *Workshop and Spring Course Dinner, Wednesday 4rd of April 2018, from 20:30 to 23:00.* The conference dinner is optional and registration is required. It will take place at the Karatello Tavern (24 Vasilissis Str. Limassol 4533 - see map at page V). Participants must bring their badge in order to attend the dinner.

GENERAL INFORMATION

Address of the venue

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

Registration

The registration will be open at the Foyer of the Mezzanine from 09:45 to 13:00 on Tuesday 3rd of April 2018 and during all the coffee breaks of the Workshop and Spring Course.

Lecture rooms

The Spring Course and the Workshop and Spring Course Keynote talks will take place at the Room Triton 1-2 located at the Mezzanine of the Poseidonia Beach Hotel. The workshop presentations will take place at the Room Triton 3 located at the Mezzanine as well.

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 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 at least ten minutes before each session. Chairs are requested to keep the sessions on schedule. Papers should be presented in the order they are listed in the programme for the convenience of attendees who may wish to go to other rooms mid-session to hear particular papers. In the case of a presenter not attending, please use the extra time for a break or a discussion so that the remaining papers stay on schedule. The PC in the lecture rooms should be used for presentations.

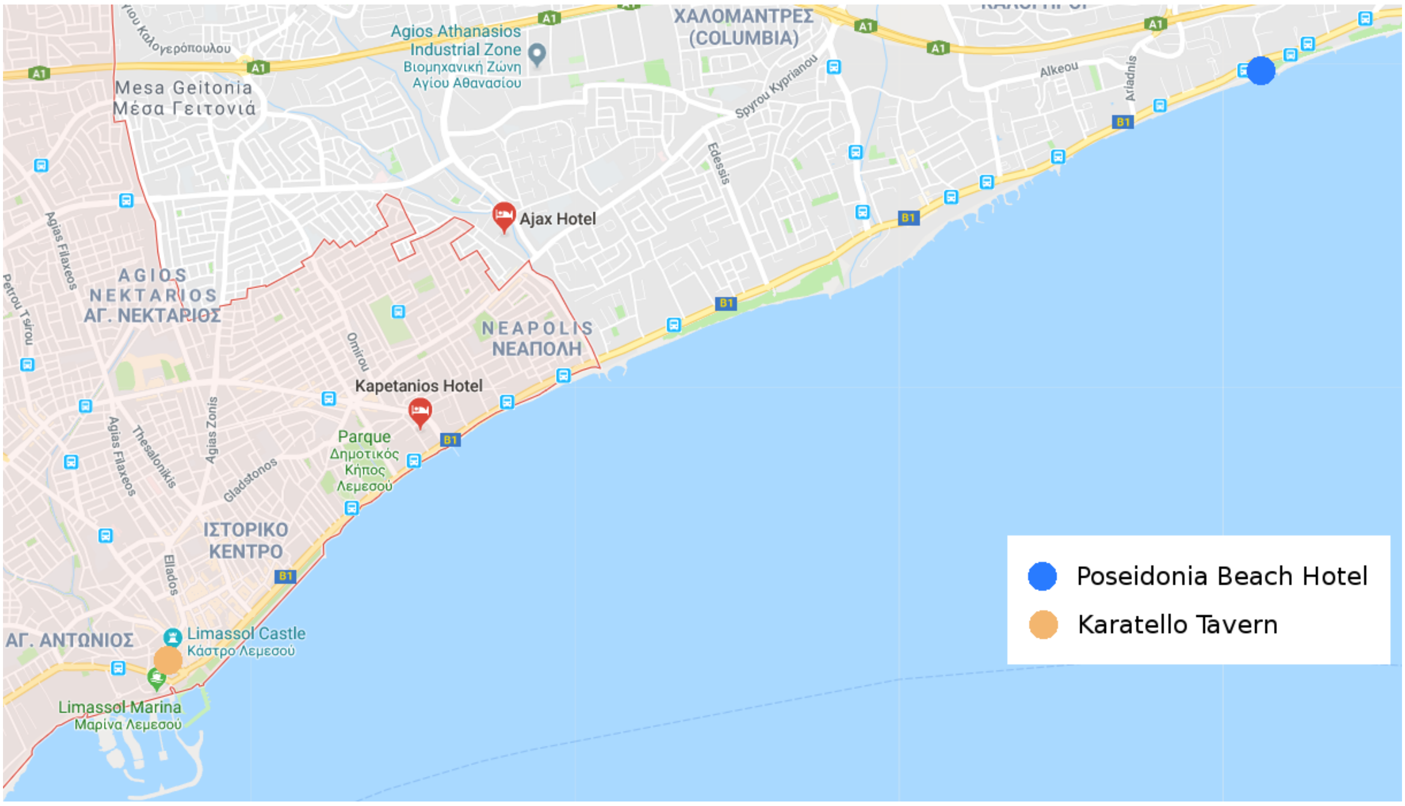
Posters

The poster sessions will take place at the Room Triton 3 located at the Mezzanine. The posters should be displayed only during their assigned session. The authors will be responsible for placing the posters in the poster panel displays and removing them after the session. The maximum size of the poster is A0.

Internet Connection

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

Map of the venue and nearby area



PUBLICATION OUTLETS

Econometrics and Statistics (EcoSta)

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

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

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

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

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

Call For Papers Econometrics and Statistics (EcoSta)

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

Papers containing novel components in econometrics and statistics are encouraged to be submitted for publication in special peer-reviewed, or regular issues of the new Elsevier journal Econometrics and Statistics (EcoSta) and its supplement Annals of Computational and Financial Econometrics. The Econometrics and Statistics (EcoSta) is inviting submissions for the special issues with deadline for submissions the 30th April 2018:

- (Part A: Econometrics) Theoretical Econometrics.
- (Part A: Econometrics) Computational Econometrics.
- (Part B: Statistics) Copulas.
- (Part B: Statistics) Neuroimaging.

Papers should be submitted using the Elsevier Electronic Submission tool EES: http://ees.elsevier.com/ecosta (in the EES please select the appropriate special issue). For further information please consult http://www.cfenetwork.org or http://www.cmstatistics.org.

Contents

Parallel Session N – CRONOSMDA2018 (Thursday 05.04.2018 at 11:30 - 13:00) 12

Parallel Session P – CRONOSMDA2018 (Thursday 05.04.2018 at 14:30 - 16:30) 13

Tuesday 03.04.2018 10:30 - 11:20 Room: Triton 1-2 Chair: Fabrizio Durante Spring Course and Workshop Keynote talk 1

Selected copula tests and models for multivariate time series

Speaker: **Ivan Kojadinovic, CNRS UMR 5142 LMA University of Pau, France**

Selected copula-based tests of stationarity and serial dependence for multivariate time series are reviewed. We then discuss a class of models for multivariate time series based on the notion of conditional copula that generalizes the well-known copula-GARCH approach frequently used for modeling financial data. We finally elaborate on tests of constancy and goodness of fit for the underlying conditional copula and discuss model selection. We provide R code illustrating the use of the presented inference and modeling procedures on real (financial) data.

Wednesday 04.04.2018 10:10 - 11:00 Room: Triton 1-2 Chair: Malgorzata Bogdan Spring Course and Workshop Keynote talk 2

Robust and consistent variable selection in high-dimensional generalized linear models

Speaker: **Elvezio Ronchetti, University of Geneva, Switzerland**

Generalized linear models are popular for modelling a large variety of data. We consider variable selection through penalized methods by focusing on resistance issues in the presence of outlying data and other deviations from assumptions. We highlight the weaknesses of widely-used penalized M-estimators, propose a robust penalized quasi-likelihood estimator, and show that it enjoys oracle properties in high dimensions and is stable in a neighborhood of the model. We illustrate its finite-sample performance on simulated and real data.

Thursday 05.04.2018 10:10 - 11:00 Room: Triton 1-2 Chair: Sara Taskinen Spring Course and Workshop Keynote talk 3

Latent variable models and pairwise likelihood framework: old and new developments

Speaker: **Irini Moustaki, London School of Economics, United Kingdom**

Latent variable models and factor models are frequently employed in social sciences where the main interest lies in measuring and relating unobserved constructs, such as emotions, attitudes, beliefs and behavior. The models become complex with the increase of the number of observed variables and the number of factors. The aim is to discuss some developments of applying a pairwise likelihood framework for the purpose of estimating the parameters of latent variable models, but also for model testing and treatment of missing values. Pairwise likelihood is a special case of composite likelihood methods that uses lower order conditional or marginal log-likelihoods instead of the full log-likelihood. Simulated and real examples will be used to illustrate the methods presented.

Thursday 05.04.2018 17:00 - 17:50 Room: Triton 1-2 Chair: Stefan Sperlich Spring Course and Workshop Keynote talk 4

Structure and estimation of VARMA and state-space systems

Speaker: **Manfred Deistler, Vienna University of Technology, Austria**

Modelling of multivariate time series by VAR, VARMA and (linear) state space systems is considered. Modelling by VAR systems is relatively simple and in a certain sense a standard task. VARMA systems (and state space systems, which are equivalent w.r.t. input-output behavior) are more flexible, but their structure is more complex, which makes estimation more involved. In particular, there is no explicit formula for the (Gaussian) maximum likelihood estimator (MLE) in this case. For the VARMA (and the state space) case we consider: 1) Properties of VARMA and linear state space systems. The relation between VARMA and state space systems; 2) Structure theory, in particular realization (i.e. construction of systems from the population second moments of the observations) and parametrization for classes of such systems; 3) Parameter estimation: MLEs, their asymptotic properties and their calculation. Initial estimators, such as subspace procedures and the Hannan Rissanen Kavalieris procedure. Approximation of MLEs, e.g. by the EM algorithm; 4) Model specification by information criteria.

Tuesday 03.04.2018 11:30 - 13:00 Parallel Session B – CRONOSMDA2018

CI025 Room Triton 1-2 SPRING COURSE SESSION I Chair: Anne Ruiz-Gazen

C0150: **Multivariate outlier detection With ICS (Part 1)**
Presenter: **Anne Ruiz-Gazen**, University Toulouse 1 Capitole, France

After a practical introduction of the general use of R for multivariate data analysis,the objective of the course is to present the Invariant Coordinate Selection (ICS) method as a tool for multivariate outlier detection. ICS was proposed in 2009 and shows remarkable properties for revealing data structures such as outliers or clusters. It is based on the simultaneous spectral decomposition of two scatter matrices and leads to an affine invariant coordinate system where the Euclidian distance corresponds to a Mahalanobis Distance (MD) in the original system. However, unlike MD, ICS makes it possible to select relevant components. This proves useful for detecting outliers lying in a small dimensional subspace for data sets in large dimensions. This context appears in particular in high reliability standards fields such as automotive, avionics or aerospace. In this context, ICS can be useful for detecting anomalies with a small proportion of false positives. The method will be illustrated on several artificial and real data sets using the recent R packages ICSOutlier and ICSShiny. The package ICSOutlier allows to choose scatter matrices, automatically select the most relevant components, calculate an outlierness index and identify potential outlying observations. The ICSShiny package provides a user-friendly application for ICS in particular for outlier detection.

CO010 Room Triton 3 HPC AND DATA ANALYSIS Chair: Aneta Karaivanova

C0178: **Studying fundamental physics using HPCs**
Presenter: **Andreas Athenodorou**, The Cyprus Institute, Cyprus

The Physics of Hadrons, the particles such as protons and neutrons which are the constituents of our world, can be studied using Lattice Quantum Chromodynamics (QCD). Quantum Chromodynamics is the theory of the strong interaction between quarks and gluons, the fundamental particles that make up composite Hadrons. Lattice QCD is a formulation that enables Physicists to study the properties of Hadrons using simulations which require Petascale even Exascale computing. We will begin by briefly explaining in an understandable manner the physics of QCD. Then, we will move to explaining Lattice QCD which is formulated on a four dimensional grid or lattice of points in space and time. QCD Fields are living on such grids. Measurements are taken on ensembles of such fields, of several TB, produced using Hybrid Monte Carlo methods which require the use of the largest available supercomputers. To this purpose Hybrid Codes which make use of MPI, OpenMP as well as GPU accelerators are being used. The solution of linear equations with a large sparse matrix is the working horse of such calculations and thus a number of Krylov Space Solvers, Parallelization strategies and several types of optimizations are applied. Followingly, we will comment on the progress of this field during the last 40 years, explain the computational as well as the data challenges of Lattice QCD, and how data science methods could assist the field. We will close by providing timely results obtained using Lattice QCD.

C0177: **ArctosPop: An integrated software package for estimating parameters of the brown bear (Ursus arctos L.) population**
Presenter: **Todor Gurov**, IICT-BAS, Bulgaria
Co-authors: Emanouil Atanassov, Aneta Karaivanova

One of the best habitats of brown bears (Ursus arctos) which are a strict protected species in Europe is located in Bulgaria. Monitoring populations of protected wildlife species is necessary for effective management and conservation of their habitats. We present the program tool ArctosPop for automatic estimation of the brown bear (Ursus arctos L.) population size in Bulgaria. The computing programme integrates statistical algorithms, which use as input data the observed data for traces of brown bears during National monitorings. As a future work, guidelines for improvement of the programme are presented.

C0175: **Study of scalability and energy efficiency of QMC algorithms on hybrid HPC systems**
Presenter: **Aneta Karaivanova**, IICT-BAS, Bulgaria
Co-authors: Emanouil Atanassov, Todor Gurov

Scalability issues and energy efficiency of some of the typical QMC algorithms for different hybrid HPC systems are studied. Depending on the features of the algorithms and the selected sequences, we obtain the optimal setup of MPI processes and OpenMP threads from point of view of speed and energy efficiency. The positive impact from using the vector instructions of the Intel Xeon Phi accelerators is demonstrated compared with CPUs and with automatic vectorization by the compiler.

Tuesday 03.04.2018 14:30 - 16:30 Parallel Session C – CRONOSMDA2018

CI027 Room Triton 1-2 SPRING COURSE SESSION II Chair: Anne Ruiz-Gazen

C0186: **Multivariate outlier detection With ICS (Part 2)**

Presenter: **Anne Ruiz-Gazen**, University Toulouse 1 Capitole, France

After a practical introduction of the general use of R for multivariate data analysis,the objective of the course is to present the Invariant Coordinate Selection (ICS) method as a tool for multivariate outlier detection. ICS was proposed in 2009 and shows remarkable properties for revealing data structures such as outliers or clusters. It is based on the simultaneous spectral decomposition of two scatter matrices and leads to an ane invariant coordinate system where the Euclidian distance corresponds to a Mahalanobis Distance (MD) in the original system. However, unlike MD, ICS makes it possible to select relevant components. This proves useful for detecting outliers lying in a small dimensional subspace for data sets in large dimensions. This context appears in particular in high reliability standards elds such as automotive, avionics or aerospace. In this context, ICS can be useful for detecting anomalies with a small proportion of false positives. The method will be illustrated on several artificial and real data sets using the recent R packages ICSOutlier and ICSShiny. The package ICSOutlier allows to choose scatter matrices, automatically select the most relevant components, calculate an outlierness index and identify potential outlying observations. The ICSShiny package provides a user-friendly application for ICS in particular for outlier detection.

CI012 Room Triton 3 CRoNoS SESSION I Chair: Ivan Kojadinovic

C0166: **A copula-based algorithm for clustering concurrent flood risks**

Presenter: **Roberta Pappada**, University of Trieste, Italy

Co-authors: Fabrizio Durante, Gianfausto Salvadori, Carlo De Michele

The impact of extreme events such as floods and droughts is often the result of compound events, arising from the joint occurrence of multiple hazards or the interaction of several variables ruling a single phenomenon. Therefore, a key issue in risk assessment framework, is the investigation of the degree of dependence between the variables at play. The focus is on flood risks, which may severely impact on the economic activities and represent a serious menace to population. In such context, copulas represent a proper tool to deal with the joint behaviour of several nonindependent random variables. A copula approach is adopted in order to develop a clustering procedure capable of detecting specific spatial sub-regions where the floods show a similar behaviour with respect to suitable (multivariate) criteria. The proposed methodology makes use of the concept of Hazard Scenarios in a copula-based setting, in order to define a similarity measure in terms of the copula expressing the dependence among floods observed at different gauge stations. A case study, involving hydrological data collected in the major Italian river basin is presented, in order to illustrate the usefulness of the proposed algorithm in the study of flood hazards.

C0170: **Copula-based clustering of time series**

Presenter: **Fabrizio Durante**, University of Salento, Italy

Clustering of time series aims to identify similarities in patterns across time. As such, several methods have been developed according to different concepts of similarity that can be based on values, functional shapes, autocorrelation structure, approximation by prototype objects, etc. We focus on a model-based clustering approach for time series by assuming that each time series follows a specific marginal model (e.g., ARMA-GARCH), while their dependence is conveniently described by means of a copula. Specifically, we review some recent clustering algorithms for time series that are based on the definition of a suitable copula-based dissimilarity measure. Such methods can conveniently describe the comovements of the time series and/or their joint tail behavior. Next, we focus on a novel clustering method for spatial time series, which performs a modified fuzzy Partitioning Around Medoids (PAM) algorithm that takes into account the copula information among time series. The different methods are presented and discussed using both simulated and real data, emphasizing their main pros and cons.

C0160: **Robust registration of probability density functions**

Presenter: **Rozenn Dahyot**, Trinity College Dublin, Ireland

Objective functions originating from optimal transport and information theory frameworks are now widely used in a range of applications, from shape registration, color transfer to machine learning. Registration of functions is also an essential processing step in functional data analysis.This talk is focusing on registering probability density functions by minimizing the robust Euclidean distance L2 or its approximation L2E. We represent probability density functions as Kernel density estimates so that the integral form with L2 has an explicit expression, and the resulting objective function is optimized efficiently with standard simulated annealing algorithms. It is shown to be robust and flexible allowing to take into account correspondences when available. For illustration, it is applied to shape registration and color transfer for image processing and computer vision applications.

C0176: **High-dimensional robust regression and outliers detection with SLOPE**

Presenter: **Malgorzata Bogdan**, University of Wroclaw, Poland

Co-authors: Stephane Gaiffas, Agathe Guilloux, Alain Virouleau

The problems of outliers detection and robust regression in a high-dimensional setting are fundamental in statistics, and have numerous applications. Following a recent set of works providing methods for simultaneous robust regression and outliers detection,we consider a model of linear regression with individual intercepts, in a high-dimensional setting. We introduce a new procedure for simultaneous estimation of the linear regression coefficients and intercepts, using two dedicated sorted-L1 penalizations, also called SLOPE.We develop a complete theory for this problem: first, we provide sharp upper bounds on the statistical estimation error of both the vector of individual intercepts and regression coefficients.Second, we give an asymptotic control on the False Discovery Rate (FDR) and statistical power for support selection of the individual intercepts. As a consequence, a procedure with guaranteed asymptotic FDR and statistical power control for outliers detection under the mean-shift model is introduce by the first time. Numerical illustrations, with a comparison to recent alternative approaches, are provided on both simulated and several real-world datasets. Experiments are conducted using an open-source software written in Python and C++.

Tuesday 03.04.2018 17:00 - 18:20 Parallel Session D – CRONOSMDA2018

CI029 Room Triton 1-2 SPRING COURSE SESSION III Chair: Anne Ruiz-Gazen

C0187: **Multivariate outlier detection with ICS (Part 3)**
Presenter: **Anne Ruiz-Gazen**, University Toulouse 1 Capitole, France

After a practical introduction of the general use of R for multivariate data analysis,the objective of the course is to present the Invariant Coordinate Selection (ICS) method as a tool for multivariate outlier detection. ICS was proposed in 2009 and shows remarkable properties for revealing data structures such as outliers or clusters. It is based on the simultaneous spectral decomposition of two scatter matrices and leads to an affine invariant coordinate system where the Euclidian distance corresponds to a Mahalanobis Distance (MD) in the original system. However, unlike MD, ICS makes it possible to select relevant components. This proves useful for detecting outliers lying in a small dimensional subspace for data sets in large dimensions. This context appears in particular in high reliability standards fields such as automotive, avionics or aerospace. In this context, ICS can be useful for detecting anomalies with a small proportion of false positives. The method will be illustrated on several artificial and real data sets using the recent R packages ICSOutlier and ICSShiny. The package ICSOutlier allows to choose scatter matrices, automatically select the most relevant components, calculate an outlierness index and identify potential outlying observations. The ICSShiny package provides a user-friendly application for ICS in particular for outlier detection.

CO008 Room Triton 3 CLUSTERING Chair: Roberta Pappada

C0163: **A new method for curves clustering in general dependence models**
Presenter: **Gianluca Sottile**, University of Palermo, Italy
Co-authors: Giada Adelfio

A new method for finding similarity of effects based on quantile regression models is proposed. Clustering of effects curves (CEC) techniques are applied to quantile regression coefficients, which are one-to-one functions of the order of the quantile. We adopt the quantile regression coefficients modeling (QRCM) framework to describe the functional form of the coefficient functions by means of parametric models. The proposed method, based on a new measure of dissimilarity that used both the shape of a curve and the distance with respect to other curves, can be utilized to cluster the effect of covariates with a univariate response variable, or to cluster a multivariate outcome, according to a variable selection perspective. The idea of combining CEC with QRCM permits simplifying computation and interpretation of the results, and may improve the ability to identify clusters. We provide computational details and cost of the algorithm, developed in the "clustEff" R package. We report simulation results in terms of number of clusters detected and two different measures of distance, comparing our approach with the existing techniques. Finally, we illustrate a variety of applications, highlighting the advantages and the usefulness of the described method.

C0169: **Model-based clustering with fixed and random covariates in R**
Presenter: **Antonio Punzo**, University of Catania, Italy
Co-authors: Angelo Mazza, Salvatore Ingrassia

Cluster-weighted models (CWMs) are mixtures of regression models with random covariates. However, besides having recently become rather popular in statistics and data mining, there is still a lack of support for CWMs within the most popular statistical suites. We introduce flexCWM, an R package specifically conceived for fitting CWMs. The package supports modeling the conditioned response variable by means of the most common distributions of the exponential family and by the t distribution. Covariates are allowed to be of a mixed-type and parsimonious modeling of multivariate normal covariates, based on the eigenvalue decomposition of the component covariance matrices, is supported. Furthermore, either the response or the covariates distributions can be omitted, yielding to mixtures of distributions and mixtures of regression models with fixed covariates, respectively. The expectation-maximization (EM) algorithm is used to obtain maximum-likelihood estimates of the parameters and likelihood-based information criteria are adopted to select the number of groups and/or the parsimonious model. For the component regression coefficients, standard errors and significance tests are also provided. Parallel computation can be used on multicore PCs and computer clusters, when several models have to be fitted.

C0153: **Agglomerative clustering with relational constraints of large datasets**
Presenter: **Vladimir Batagelj**, IMFM, Slovenia

Agglomerative clustering algorithms for solving clustering problems with relational constraints were proposed already in eighties. A problem with these algorithms is their scalability. Because they are based on a dissimilarity matrix they can be applied to data sets with up to some ten thousands of units. We discuss two approaches for agglomerative clustering of large datasets. Both are based on the idea to compute the dissimilarities only between the related (with constraints) units and the assumption that the constraints network is sparse. The first approach is based on the introduction of new dissimilarities between clusters, the second on "classical" dissimilarities between cluster representatives. Both approaches were implemented in R and will be illustrated on some real-life datasets.

Tuesday 03.04.2018 18:30 - 19:30 Parallel Session E – CRONOSMDA2018

CI031 Room Triton 1-2 SPRING COURSE SESSION IV Chair: Anne Ruiz-Gazen

C0188: **Multivariate outlier detection With ICS (Part 4)**
Presenter: **Anne Ruiz-Gazen**, University Toulouse 1 Capitole, France

After a practical introduction of the general use of R for multivariate data analysis,the objective of the course is to present the Invariant Coordinate Selection (ICS) method as a tool for multivariate outlier detection. ICS was proposed in 2009 and shows remarkable properties for revealing data structures such as outliers or clusters. It is based on the simultaneous spectral decomposition of two scatter matrices and leads to an affine invariant coordinate system where the Euclidian distance corresponds to a Mahalanobis Distance (MD) in the original system. However, unlike MD, ICS makes it possible to select relevant components. This proves useful for detecting outliers lying in a small dimensional subspace for data sets in large dimensions. This context appears in particular in high reliability standards fields such as automotive, avionics or aerospace. In this context, ICS can be useful for detecting anomalies with a small proportion of false positives. The method will be illustrated on several artificial and real data sets using the recent R packages ICSOutlier and ICSShiny. The package ICSOutlier allows to choose scatter matrices, automatically select the most relevant components, calculate an outlierness index and identify potential outlying observations. The ICSShiny package provides a user-friendly application for ICS in particular for outlier detection.

CP001 Room Triton 3 POSTER SESSION Chair: Maria Elena Fernandez Iglesias

C0154: **Resampling approaches for multivariate data: Theory, R-package and applications**
Presenter: **Sarah Friedrich**, University of Ulm, Germany
Co-authors: Markus Pauly

In many experiments in the life sciences several endpoints, potentially measured on different scales, are recorded per subject. Classical MANOVA models assume normally distributed errors and homogeneity of the covariance matrices, two assumptions that are often not met in practice. Inference is further complicated, if covariance matrices are singular. We propose a test statistic for factorial MANOVA designs which incorporates general heteroscedastic models with possibly singular covariance matrices. Different bootstrap techniques are used in order to derive inference even for small samples. The methods are implemented in the R package MANOVA.RM. The package allows for different factorial designs with nested and crossed factors and comes with a plotting routine and a graphical user interface. We present the main functionalities of the package and use it to analyse a practical data set.

C0173: **Periodic bivariate count time series models**
Presenter: **Magda Monteiro**, University of Aveiro, Portugal
Co-authors: Isabel Pereira, Manuel Scotto

In real life, there are several count time series that exhibit periodicity and are also related to each other allowing there joint modelling as a bivariate time series. Examples can be found in different field areas such as environmental (monthly number of fires in two neighbour counties), tourism (monthly number of guests from neighbours hotels), labour market (monthly number of short and long-term unemployed in a county) among others. As in other types of time series, the modelling of these time series can be done using different approaches of which we highlight models based on thinned operations and bivariate dynamic factor models. We will present a comparative study of a periodic bivariate integer-value autoregressive (PBINAR) model and a bivariate dynamic factor model under the context of forest fires application.

C0174: **Factor model estimation by composite minimization**
Presenter: **Matteo Farne**, University of Bologna, Italy
Co-authors: Angela Montanari

The problem of factor model estimation in large dimensions is addressed under the low rank plus sparse assumption. Existing approaches based on PCA like POET estimator fail to catch low rank spaces characterized by non-spiked eigenvalues, as in this case the asymptotic consistency of PCA defaults. UNALCE, an alternative approach based on the minimization of a low rank plus sparse decomposition problem, is shown to produce the covariance estimate with the least possible dispersed eigenvalues among all the matrices having the same rank of the low rank component and the same support of the sparse component. Consequently, if dimension and sample size are fixed, loadings and factor scores estimated via UNALCE provide the tightest possible error bound. The result is based on the sample eigenvalue dispersion lemma. The effectiveness of UNALCE factor estimates is finally explored in an exhaustive simulation study, which clarifies that the gain of UNALCE is larger as the latent eigenvalues are less spiked and the sparse component is more sparse.

C0179: **Impact of rainfall on the greater flow peaks in Esva river Basin in NW of Spain and relations to land use changes**
Presenter: **Maria Elena Fernandez Iglesias**, University of Oviedo, Spain
Co-authors: Gil Gonzalez-Rodriguez, Jorge Marquinez, Maria Fernandez-Garcia

The river channel in the Esva basin, a coastal catchment of the Cantabrian region (Northwest of Spain), has experienced a slight active channel-width decrease from 1957 to 1985 (<1%) and more important decrease (close to 13%) from 1985 to 2003. This trend is well related to the main changes in the forest cover, which also increases slightly from 1957 to 1985 and more importantly after 1985 until now. Models for different scenarios were applied for the daily datasets before and after 1985 with the aim of identifying whether the changes in the land uses might influence in the hydrological response of the river to the rainfall. The preliminary results focused on flow peak events conclude a light increase in this time reaction rainfall-flow in agreement with the land usage changes.

Wednesday 04.04.2018 09:00 - 10:00 Parallel Session F – CRONOSMDA2018

CI033 Room Triton 1-2 SPRING COURSE SESSION V Chair: Anne Ruiz-Gazen

C0196: **Multivariate outlier detection with ICS (Part 5)**

Presenter: **Anne Ruiz-Gazen**, University Toulouse 1 Capitole, France

After a practical introduction of the general use of R for multivariate data analysis,the objective of the course is to present the Invariant Coordinate Selection (ICS) method as a tool for multivariate outlier detection. ICS was proposed in 2009 and shows remarkable properties for revealing data structures such as outliers or clusters. It is based on the simultaneous spectral decomposition of two scatter matrices and leads to an affine invariant coordinate system where the Euclidian distance corresponds to a Mahalanobis Distance (MD) in the original system. However, unlike MD, ICS makes it possible to select relevant components. This proves useful for detecting outliers lying in a small dimensional subspace for data sets in large dimensions. This context appears in particular in high reliability standards fields such as automotive, avionics or aerospace. In this context, ICS can be useful for detecting anomalies with a small proportion of false positives. The method will be illustrated on several artificial and real data sets using the recent R packages ICSOutlier and ICSShiny. The package ICSOutlier allows to choose scatter matrices, automatically select the most relevant components, calculate an outlierness index and identify potential outlying observations. The ICSShiny package provides a user-friendly application for ICS in particular for outlier detection.

Wednesday 04.04.2018 11:30 - 13:00 Parallel Session H – CRONOSMDA2018

CI035 Room Triton 1-2 SPRING COURSE SESSION VI Chair: Simon Caton

C0183: **Parallelisation of R models with h2o (Part 1)**
Presenter: **Simon Caton**, National College of Ireland, Ireland

Model scaling is becoming increasingly necessary as datasets increase in size, but also to facilitate core aspects of model building, model prototyping, and model selection. In this session, we will explore the application of h2o to facilitate the parallelisation of R models. The session will begin with parallelising a selection of multivariate methods to use multiple cores on participants' machines. From here, it will move towards leveraging cloud resources to further increase model scalability and correspondingly reduce runtimes. It will culminate with advice on appropriate uses of cloud and other parallel architectures for model building.

CI002 Room Triton 3 CRONOS SESSION II Chair: Roland Fried

C0195: **Modeling heterogeneity by structural varying coefficients models in presence of endogeneity**
Presenter: **Stefan Sperlich**, Univserity of Geneva, Switzerland

High degrees of heterogeneity across the studied units can make the estimates of structural parameters inaccurate and uninformative. First, we reviews how semiparametric varying coefficient models can identify cross-sectional heterogeneity. Next, we discuss how different potential sources of endogeneity problems can be addressed. It is explained thereby why modelling coefficient variation across groups increases the credibility and interpretability of the assumptions needed when using instrumental variable estimation. They are therefore not just a nice compromise between non- and fully parametric models to circumvent the curse of dimensionality, they are also a good compromise between these two extremes regarding causal inference. This will be illustrated along various models taken from applied econometrics. In order to make the econometric specification theory-consistent, all considerations and developments are embedded in the modeling of heterogeneous structural equations rather than in purely theoretical remedies.

C0180: **Maximizing the discriminatory power of bankruptcy prediction models using different merit functions**
Presenter: **Christakis Charalambous**, University of Cyprus, Cyprus
Co-authors: Spiros Martzoukos, Zenon Taoushianis

Acknowledging the economic benefits associated with the development of powerful bankruptcy prediction models, this paper presents a methodology for maximizing their discriminatory power, when considering both probabilistic and linear response functions for the output of the models as well as different merit functions, used to obtain the coefficient estimates. For our analysis, we use accounting and market-related information for a sample of U.S. public bankrupt and healthy firms between 1990 and 2015. Results show an improvement in the discriminatory power when we implement our approach as compared with traditional approaches, such as logistic regression models. We also find that using models with a merit function that accounts for outliers, yields better performance. Among all models, the neural network is the best performing model. More importantly, results hold under different tests.

C0158: **The role of generalized means in multivariate extreme value statistics**
Presenter: **Ivette Gomes**, FCiencias.ID, Universidade de Lisboa and CEAUL, Portugal

Modern risk assessment of the amount of tail dependence is nowadays crucial in the most diverse fields, like finance and insurance, among others, and the correlation structure is usually not enough to describe such a tail dependence. Given a pair (X,Y), with margins, (F_X,F_Y), the *tail dependence coefficient* (TDC), denoted by η, can be defined as a limiting conditional probability, $\eta := \lim_{t\to\infty} P\left(F_X(X) > 1-1/t | F_Y(Y) > 1-1/t\right)$. The standardization of the margins to unit Fréchet margins, enables the estimation of the TDC in a way similar to the estimation of the *extreme value index* (EVI) associated with $Z := \min(X,Y)$. Indeed, $P(Z>z) = P(X>z, Y>z) = z^{-1/\eta}\mathcal{L}(z)$, where $\mathcal{L}(\cdot)$ is a slowly-varying function at infinity. *Generalized means* (GMs) have recently been successfully used for the estimation of parameters of extreme events, like the EVI and the value at risk. Now, GMs are used under a multivariate framework, essentially for the estimation of the TDC, in bivariate extreme value statistics. Associated asymptotically unbiased estimators are also constructed. The finite-sample behavior as well as robustness, regarding sensitivity to the extreme value dependence assumption, is assessed through small-scale Monte-Carlo simulation studies.

Wednesday 04.04.2018 14:30 - 16:30 Parallel Session I – CRONOSMDA2018

CI037 Room Triton 1-2 SPRING COURSE SESSION VII Chair: Simon Caton

C0189: **Parallelisation of R models with h2o (Part 2)**
Presenter: **Simon Caton**, National College of Ireland, Ireland

Model scaling is becoming increasingly necessary as datasets increase in size, but also to facilitate core aspects of model building, model prototyping, and model selection. In this session, we will explore the application of h2o to facilitate the parallelisation of R models. The session will begin with parallelising a selection of multivariate methods to use multiple cores on participants' machines. From here, it will move towards leveraging cloud resources to further increase model scalability and correspondingly reduce runtimes. It will culminate with advice on appropriate uses of cloud and other parallel architectures for model building.

CG015 Room Triton 3 CONTRIBUTIONS TO MULTIVARIATE DATA ANALYSIS Chair: Stefan Sperlich

C0155: **Paragraph vector topic modeling: Methods and applications to identify and measure the diffusion of innovations**
Presenter: **David Lenz**, Justus-Liebig University Giessen, Germany
Co-authors: Peter Winker

Topic modeling became an intensively researched area lately, mainly due to the vast availability of written information available through the world wide web and the improvements in methods to analyze these datasets. In natural language processing (NLP), topic modeling describes a set of methods to extract the latent topics that occur in a collection of documents. Several new methods have recently been proposed to improve the topic generation process, however examination of the generated topics is still mostly based on unsatisfactory practices. Typically, the generated topics are identified by looking at a list of top words, which requires significant mental effort and can be tedious and demanding work. Our contribution is threefold: 1) We present a novel topic modeling approach based on neural embeddings and Gaussian mixture modeling, which is shown to generate coherent and meaningful topics. 2) We propose a topic report style sheet based on dimensionality reduction techniques and model generated document vector features which helps to easily identify topics and significantly reduces the required mental overhead. 3) Lastly, we demonstrate on a newsticker corpus how our approach can be used to measure the diffusion of innovations.

C0164: **Constructing leading economic indicators for the Philippineeconomy**
Presenter: **Dennis Mapa**, University of the Philippines Diliman, Philippines

The aim is to propose three models (a Dynamic Factor model, a Hybrid Dynamic Factor-Vector AutoRegressive (DF-VAR) model and a Dynamic Factor-Mixed Frequency (DF-MF) model) in nowcasting the movements and growth rates of the country's quarterly Gross Domestic Product (GDP) using monthly indicator variables. The DF, DF-VAR and DF-MF are alternative models to the usual time series econometric models used in forecasting GDP growth rates utilizing temporal aggregation. The idea behind the DF model is the stylized fact that economic movements evolve in a cycle and are correlated with co-movements in a large number of economic series. The DF model is a commonly used data reduction procedure that assumes economic shocks driving economic activity arise from unobserved components or factors. The DF model aims to parsimoniously summarize information from a large number of economic series to a small number of unobserved factors. The DF model assumes that co-movements of economic series can be captured using these unobserved common factors. While the DF model captures the movements in the GDP growth, combining the DF with the Vector AutoRegressive (VAR) model (or with the Mixed Frequency model) will be useful is also nowcasting the GDP growth rates and not just the movements.

C0156: **Domain selection for multivariate functional data classification**
Presenter: **Nicolas Hernandez**, Universidad Carlos III de Madrid, Spain
Co-authors: Gabriel Martos, Alberto Munoz

A domain selection approach is proposed for classification problems in functional data. Consider two samples of random elements $f_1, \dots, f_n$ and $g_1, \dots, g_m$ in $L^2(X)$ generated from the functional stochastic models $f_i(x) = \mu_k(x) + \varepsilon_i(x)$ for $i = 1, \dots, n$ and $g_j(x) = \mu_k(x) + \varepsilon_j(x)$ for $j = 1 \dots, m$ respectively and defined on the same domain $X = [0,1]$. The function $\mu_k(x)$ is the mean function for $k = f, g$ and $\varepsilon(x)$ is a random and independent functional error that captures the variability within each class. In this setting, we propose to use a local–inner product parametrized by the vector $\theta = (\theta_1, \theta_2)$, with $0 \leqslant \theta_1 < \theta_2 \leqslant 1$, such that, $\langle f, g \rangle_\theta = \int_{\theta_1}^{\theta_2} f(x)g(x)dx$. The proposed inner–product induce a local–metric in the space of random elements $L^2(X)$. The optimization of θ is presented as a domain selection technique, where the optimization goal pursue the minimization of the misclassification error rate when classifying samples of random functions.

C0182: **A new version of I^2 with emphasis on diagnostic problems**
Presenter: **Heinz Holling**, University of Munster, Germany

A common measure for heterogeneity in meta-analysis is Higgins' I^2. This measure has been criticized for being confounded by the study-specific sample size, in the sense that different I^2-values can be achieved for the same value of across-study variance if only the study-specific variance is varying enough. In particular, I^2 approaches one for any value of the heterogeneity variance (variance across studies) if the within-study variance becomes large. It will be shown that the within-study variance is asymptotically identical to the harmonic mean of the study-specific variances and, for any number of studies, is at least as large as the harmonic mean with the inequality being sharp if all study-specific variances agree. Then, a new measure, which is unconfounded by sample size is proposed. A detailed simulation study has been launched and the results indicate that the newly suggested measure of heterogeneity has beneficial statistical properties. Furthermore, Higgins' I^2 and the new measure are exemplified at hand of some meta-analytic case studies.

Wednesday 04.04.2018 17:00 - 18:20 Parallel Session J – CRONOSMDA2018

CI039 Room Triton 1-2 SPRING COURSE SESSION VIII Chair: Simon Caton

C0190: **Parallelisation of R models with h2o (Part 3)**
Presenter: **Simon Caton**, National College of Ireland, Ireland

Model scaling is becoming increasingly necessary as datasets increase in size, but also to facilitate core aspects of model building, model prototyping, and model selection. In this session, we will explore the application of h2o to facilitate the parallelisation of R models. The session will begin with parallelising a selection of multivariate methods to use multiple cores on participants' machines. From here, it will move towards leveraging cloud resources to further increase model scalability and correspondingly reduce runtimes. It will culminate with advice on appropriate uses of cloud and other parallel architectures for model building.

CO014 Room Triton 3 ANALYSIS OF HIGH DIMENSIONAL COMPLEX DATA Chair: Andreas Artemiou

C0152: **Projection-based extension of estimators for tensor data**
Presenter: **Joni Virta**, Aalto University, Finland
Co-authors: Klaus Nordhausen

A general approach for extending multivariate estimators to matrix- and tensor-valued data is proposed. The extension is based on using random projections to project out all but one of the dimensions of a tensor and compute the multivariate estimator for each projection. The mean of the obtained set of estimates is used as the final, joint estimate. In some basic cases the resulting estimator can be given a closed form and particular ones are shown to coincide with existing methodology. Comparisons with competing methods show that the extensions prove useful in extracting components for classification and yield an efficient estimator for sufficient dimension reduction.

C0159: **Mapping resting state functional brain connectivity with high-dimensional penalized graphical models**
Presenter: **Eugen Pircalabelu**, KU Leuven; BE 0419.052.173, Belgium
Co-authors: Gerda Claeskens, Lourens Waldorp

Brain networks from fMRI datasets that do not all contain measurements on the same set of regions are estimated. For certain datasets, some of the regions have been split in smaller subregions, while others have not been split. This gives rise to the framework of mixed scale measurements and the purpose is to estimate sparse graphical models.The resulting graphical models combine information from several subjects, overcome the problem of having data on different coarseness levels and take into account that dependencies exist between a coarse scale node and its finer scale nodes, since finer scale nodes are obtained by splitting coarser ones. Our procedure is directed towards estimating effects between split and unsplit regions, since this offers insight into whether a certain large ROI is constructed by aggregating homogeneous or heterogeneous parts of the brain. The method results in estimating graphical models for each coarseness level in the analysis and identifies possible connections between a large region and its subregions, referred to as between level edges. We also investigate zooming-in and out procedures to assess the evolution of edges across the coarseness scales.

C0171: **Copula Gaussian graphical models for functional data**
Presenter: **Eftychia Solea**, University of Cyprus, Cyprus

The problem of constructing statistical graphical models for functional data is considered; that is, the observations on the vertices are random functions. This types of data are common in medical applications such as EEG and fMRI. Recently published functional graphical models rely on the assumption that the random functions are Hilbert-space-valued Gaussian random elements. We relax this assumption by introducing a copula Gaussian random elements Hilbert spaces, leading to what we call the Functional Copula Gaussian Graphical Model (FCGGM). This model removes the marginal Gaussian assumption but retains the simplicity of the Gaussian dependence structure, which is particularly attractive for large data. We develop four estimators, together with their implementation algorithms, for the FCGGM. We establish the consistency and the convergence rates of one of the estimators under different sets of sufficient conditions with varying strengths. We compare our FCGGM with the existing functional Gaussian graphical model by simulation, under both non-Gaussian and Gaussian graphical models, and apply our method to an EEG data set to construct brain networks.

Wednesday 04.04.2018 18:30 - 19:30 Parallel Session K – CRONOSMDA2018

CI041 Room Triton 1-2 SPRING COURSE SESSION IX Chair: Simon Caton

C0191: **Parallelisation of R models with h2o (Part 4)**

Presenter: **Simon Caton**, National College of Ireland, Ireland

Model scaling is becoming increasingly necessary as datasets increase in size, but also to facilitate core aspects of model building, model prototyping, and model selection. In this session, we will explore the application of h2o to facilitate the parallelisation of R models. The session will begin with parallelising a selection of multivariate methods to use multiple cores on participants' machines. From here, it will move towards leveraging cloud resources to further increase model scalability and correspondingly reduce runtimes. It will culminate with advice on appropriate uses of cloud and other parallel architectures for model building.

Thursday 05.04.2018 09:00 - 10:00 Parallel Session L – CRONOSMDA2018

CI043 Room Triton 1-2 SPRING COURSE SESSION X **Chair: Simon Caton**

C0197: **Parallelisation of R Models with h2o (Part 5)**

Presenter: **Simon Caton**, National College of Ireland, Ireland

Model scaling is becoming increasingly necessary as datasets increase in size, but also to facilitate core aspects of model building, model prototyping, and model selection. In this session, we will explore the application of h2o to facilitate the parallelisation of R models. The session will begin with parallelising a selection of multivariate methods to use multiple cores on participants' machines. From here, it will move towards leveraging cloud resources to further increase model scalability and correspondingly reduce runtimes. It will culminate with advice on appropriate uses of cloud and other parallel architectures for model building.

Thursday 05.04.2018 11:30 - 13:00 Parallel Session N – CRONOSMDA2018

CI045 Room Triton 1-2 SPRING COURSE SESSION XI Chair: Roland Fried

C0184: **robts - an R-package for robust time series and changepoint analysis**
Presenter: **Roland Fried**, TU Dortmund University, Germany

The progress on our R-package robts is reported, which is available from R-Forge. Our package works under the assumption of short range dependence and provides different techniques for robust estimation of autocorrelations, partial autocorrelations and spectral densities, for robust fitting of autoregressive time series models, for model diagnostics and prediction. Since many time series models assume second order stationarity, we include robust tests for checking the stationarity of the mean, the variance and the autocovariances. Extensions to multivariate time series analysis are a task for future work.

CI004 Room Triton 3 CRONOS SESSION III Chair: Irini Moustaki

C0161: **Weighting of parts in compositional data analysis with applications**
Presenter: **Karel Hron**, Palacky University, Czech Republic
Co-authors: Juan Jose Egozcue, Vera Pawlowsky-Glahn, Javier Palarea-Albaladejo, Peter Filzmoser, Alessandra Menafoglio

Statistical analysis of D-part compositional data, i.e. multivariate observations carrying relative information, using tools derived from the Aitchison geometry on the simplex assumes uniform distribution as reference measure of the measurable space. We consider its decomposition into D categories corresponding to the compositional parts. A change of the reference measure corresponds to weighting of parts, which impacts on the algebraic-geometrical structure on the simplex and the coordinate representation of compositional data. This latter is a requirement for the use of standard tools of multivariate statistics in compositional analysis. The choice of weights depends on the nature of the problem, and might reflect, e.g., measurement imprecision of compositional parts, presence of values below detection limit, and so on. Specifically, it is possible to show that if the weights of selected parts approach zero then the space of the respective subcomposition is obtained as the limiting case. Theoretical outputs will be demonstrated on simulated and real data.

C0165: **On the coordinate-free analysis of multivariate categorical data**
Presenter: **Tamas Rudas**, Hungarian Academy of Sciences Centre for Social Sciences, Hungary

In multivariate statistics, the sample space of is usually the Cartesian product of the ranges of the variables involved. Even the simplest structures used in statistics, like independence, rely heavily on this structure. Basic concepts of statistical modeling are introduced, which can be applied when the structure of the sample space is different. First, motivating examples are presented, then coordinate free exponential families of probability distributions are introduced, which postulate simple multiplicative structures. Some of the properties of these families are similar to that of log-linear models, but the maximum likelihood estimates under these models have a few very surprising characteristics.

C0172: **Generalized linear latent variable models in the analysis of multivariate abundance data**
Presenter: **Sara Taskinen**, University of Jyvaskyla, Finland

In many ecological studies, counts or biomass of interacting species are collected from several sites. Such data are often very sparse, high-dimensional and include highly correlated responses, and the main aim of the statistical analysis is to understand relationships among such multiple, correlated responses. We show how generalized linear latent variable models can be used to analyze data common in ecological studies. By extending the standard generalized linear modelling framework to include latent variables, we can account for any covariation between species not accounted for by the predictors, species interactions and correlations driven by missing covariates. Fast and efficient maximum likelihood based algorithms for fitting the models will be discussed and simulations are used to study the finite-sample properties of the resulting estimates. It is shown that especially the variational approximation method performs well when fitting GLLVMs. We will also illustrate how GLLVMs can be used when developing new tools for ecological monitoring of Finnish peatlands.

Thursday 05.04.2018 14:30 - 16:30 Parallel Session P – CRONOSMDA2018

CI047 Room Triton 1-2 SPRING COURSE SESSION XII Chair: Cristian Gatu

C0185: **Computational strategies for regression model selection**
Presenter: **Cristian Gatu**, Alexandru Ioan Cuza University of Iasi, Romania

Computational strategies for computing the best-subset regression models are proposed. The algorithms are based on a regression tree structure that generates all possible subset models. An efficient branch-and-bound algorithm that finds the best submodels without generating the entire tree is described. Specifically, the computational burden is reduced by pruning the non-optimal subtrees. Strategies and approximate algorithms that improve the computational performance are investigated. Further, this strategies are adapted to solve the problem of regression subset selection under the condition of non-negative coefficients. The solution is based on an alternative approach to quadratic programming that derives the non-negative least squares by solving the normal equations for a number of unrestricted least squares subproblems. This innovative approach is computationally superior to the straight-forward method that would estimate the corresponding non-negative least squares of all possible submodels in order to select the best one. The R package "lmSubsets" for regression subset selection is introduced and described. The package aims to provide a versatile tool for subset regression.

CI006 Room Triton 3 CRONOS SESSION IV Chair: Gil Gonzalez-Rodriguez

C0168: **Generic software tools for model checking and display**
Presenter: **John Hinde**, NUI Galway, Ireland
Co-authors: Rafael de Andrade Moral, Amirhossein Jalali

Open systems, like R, allow the rapid dissemination of new methodology through user-contributed packages. There has been an explosion of these, however, the focus is typically on new models, methods of analysis, etc, rather than on generic tools that may enhance other aspects of the modelling process. We consider from the point of view of both model checking and model display/interrogation. We illustrate the first with the hnp package for model checking using half-normal plots of diagnostic quantities, eg residuals. This package can be used with many common univariate response model objects and can easily be extended to any new class. We also consider the extension of these ideas to bivariate response models. Nomograms (graphical tools for complex expressions) have received renewed interest in recent years as a form of visualisation tool. The DynNom package is an example of a Shiny app for model display using dynamic nomograms. We illustrate its use and discuss the general approach and its role as a communication and dissemination tool.

C0162: **Detection and recovery from inconsistencies in the general linear model with singular dispersion matrix**
Presenter: **Marc Hofmann**, University of Oviedo, Spain
Co-authors: Ana Colubi, Erricos John Kontoghiorghes

A new method to recover from an insconsistent GLM is proposed. The GLM is reformulated as a GLLSP. The minimal set of observations thatexplain the inconsistencies in the model can be identified by solving a combinatorial sparse approximation problem. An exhaustive algorithm isproposed. Gram-Schmidt orthogonalization is used as the main computational tool. When the number of observations is large, non-exhaustivealgorithms can be employed instead.

C0167: **M.J.D. Powell's software packages for constrained optimization calculations**
Presenter: **Ioannis Demetriou**, University of Athens, Greece

Michael J. D. Powell (1936 - 2015) has been a world's leader in optimization. An integral part of his research work has been software development. A brief survey is given of his software for constrained optimization calculations. Attention is given to the purpose of each software package instead of to its details. All the packages to be mentioned are important to Multivariate Data Analysis.

Authors Index