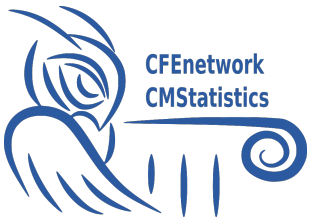
PROGRAMME AND ABSTRACTS

25th International Conference on

Computational Statistics (COMPSTAT 2023)

http://www.compstat2023.org

Birkbeck, University of London, UK
22-25 August 2023



ISBN: 9789073592414

COMPSTAT 2023 Scientific Program Committee:

Ex-officio:
COMPSTAT 2023 organiser and chairperson of the SPC: Erricos Kontoghiorghes.
Past COMPSTAT organiser: Alessandra Luati and Maria Brigida Ferraro.
Next COMPSTAT organiser: Ana Colubi and Peter Winker.
Incoming IASC-ERS Chairman: M. Brigida Ferraro.

Members:
Eva Cantoni, Michele Guindani, Ivan Kojadinovic, Ioanna Manolopoulou, Mike So and Sara Taskinen.

Consultative Members:
Representative of the IFCS: Angela Montanari.
Representative of the ARS of IASC: Ray-Bing Chen.
Representative of CMStatistics: Xuming He.

Local Organizing Committee:
Liudas Giraitis, Janet Godolphin, George Kapetanios, Stephen Pollock, George Roussos, F. Javier Rubio and David Weston.

Dear Colleagues and Friends,

Welcome to the 25th International Conference on Computational Statistics (COMPSTAT 2023) in London. This special edition of COMPSTAT in an odd year is a result of the necessary postponements due to the pandemic. Thankfully, we have been able to proceed with the activities planned earlier, and we truly appreciate your support throughout this period.

The organization of this event has been primarily led by members of Birkbeck, University of London, with the assistance of esteemed international researchers. COMPSTAT, initiated by the European Regional Section of the International Association for Statistical Computing (IASC-ERS), a society of the International Statistical Institute (ISI), holds a reputable position as one of the most esteemed global conferences in Computational Statistics, regularly attracting numerous researchers and practitioners.

Since its inception in 1974 in Vienna, COMPSTAT has gained recognition as an ideal platform for presenting exceptional theoretical and applied work, fostering interdisciplinary research, and facilitating connections among researchers with shared interests.

The conference program includes 30 contributed sessions, 6 invited sessions, 3 keynote talks, 25 organized sessions, and 2 tutorials, with approximately 290 participants. To accommodate various preferences, COMPSTAT 2023 will be conducted in a hybrid format, with all sessions live-streamed, offering participants the option to attend the conference online.

We would like to express our heartfelt appreciation to all the authors and participants who have contributed to the success of COMPSTAT 2023. We are sincerely grateful to our sponsors, the scientific program committee, session organizers, local hosts, and the many volunteers whose efforts have played a crucial role in making this conference possible. We acknowledge and commend their dedication and support.

As we look forward to the future, we extend a warm invitation to all of you to join us in Giessen, Germany, from 27th to 30th August 2024, for the 26th edition of COMPSTAT. Our best wishes for success go to the chairs of the upcoming edition.

Once again, we thank each one of you for your enthusiastic participation and eagerly anticipate meeting you all in London for an intellectually stimulating and memorable experience.

Erricos Kontoghiorghes
Organiser and chairperson

SCHEDULE COMPSTAT 2023

2023-08-22	2023-08-23	2023-08-24	2023-08-25
A - Keynote 09:00 - 10:00	**F** 09:00 - 10:30	**I** 09:00 - 10:00	**M** 09:00 - 10:00
Coffee break 10:00 - 10:30	**Coffee break** 10:30 - 11:00	**Coffee break** 10:00 - 10:30	**Coffee break** 10:00 - 10:30
B 10:30 - 12:30	**G** 11:00 - 12:30	**J** 10:30 - 12:30	**N** 10:30 - 12:00
Lunch break 12:30 - 14:15	**Lunch break** 12:30 - 14:15	**Lunch break** 12:30 - 14:15	**O - Keynote** 12:10 - 13:10
C 14:15 - 15:45	**H** 14:15 - 15:45	**K** 14:15 - 15:45	
Coffee break 15:45 - 16:15		**Coffee break** 15:45 - 16:15	
D 16:15 - 17:45		**L** 16:15 - 17:45	
E - Keynote 17:55 - 18:55			
Welcome reception 19:00 - 20:30		**Conference dinner** 19:00 - 21:30	

Tutorials, social events and general information

TUTORIALS - COMPSTAT 2023

The first tutorial is virtual, given by Prof. Alessandra Luati (*Dynamic models for multiple quantiles*), Tuesday 22.8.2023, 10:30 - 12:30. The second is given by Prof. Francisco Javier Rubio (*Bayesian variable selection for survival data: Theory, methods, software and applications*), Thursday, 24.8.2023, 10:30 - 12:30.

SOCIAL EVENTS - COMPSTAT 2023

- *The coffee breaks* will take place in Room 305, Floor 3, Birkbeck Central Building. You must have your conference badge in order to attend the coffee breaks.
- *Welcome Reception, Tuesday, 22nd August 2023, 19:00-20:30.* The Welcome Reception will take place in Room 305, Floor 3, Birkbeck Cental Building, and is open to all registrants who had preregistered and accompanying persons who purchased a reception ticket. Participants must bring their conference badges in order to attend the reception. Preregistration is required due to health and safety reasons.
- *Afternoon tea, Wednesday, 23rd August 2023, 16:00-17:30.* Traditional full afternoon tea at the Ambassadors Bloomsbury Hotel, 12 Upper Woburn Place, London, WC1H 0HX, United Kingdom. This event is optional, and registration is required. Participants must bring their conference badge in order to attend the event. Information about the booking is embedded in the QR code on the conference badge.
- *Conference Dinner, Thursday, 24th August 2023, 20:00-22:30.* The conference dinner will take place at the Ambassadors Bloomsbury Hotel. The conference dinner is optional, and registration is required. Participants must bring their conference badges to attend the conference dinner. Information about the purchased conference dinner ticket is embedded in the QR code on the conference badge.

Address of venues

The Conference venue is the Birkbeck Central building, Birkbeck University of London, UK, Malet St, London WC1E 7HY.

Registration

Registration will be open on Tuesday from 08:00 to 18:00, Wednesday from 08:40 to 15:00, Thursday from 08:40 to 17:00 and Friday from 08:40 to 12:30. It will take place in the ground-level Hall of the Birkbeck Central building

Presentation instructions

The opening and the keynote on Monday will take place in CLO B01, on the basement of the Clore Management Center. The rest of the conference will run in the Birkbeck Central building, mainly on Floor 3. BCB 206 on Floor 2 can be used as a quiet room to attend virtual sessions from their own devices and will also be used for the closing Keynote talk. The poster sessions will take place online, but in-person participants are invited to meet in the registration area with their own laptops for related discussions. The virtual presentations will take place through Zoom. Speakers should have a stable internet connection, and ensure their video and audio are working. They will share their slides when the chair requires it, present their talk, and answer questions after the presentation. The in-person speakers must share presentations through the Zoom session open on the desktop in the conference rooms. The rooms have a webcam, and an omnidirectional desk microphone that collects the sound around the PC desk to make the live streaming easy. Detailed indications for speakers in either virtual or hybrid sessions can be found on the website. As a general rule, each speaker has 20 minutes, including 2-3 minutes for discussion. Strict timing must be observed.

Posters

The poster sessions will take place through Gather Town. The posters should be sent in png format to info@compstat2023.org by 15th August. Landscape orientation is advisable. Detailed indications for the poster presentations can be found on the website.

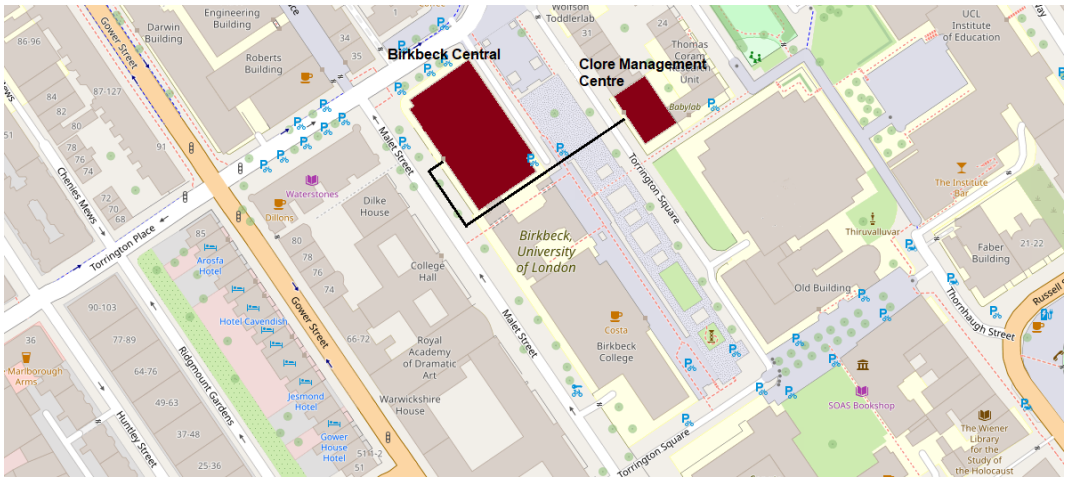
Session chairs

The session chairs will be responsible for introducing the session, the speakers and coordinating the discussion time. A member of the conference staff, identified on Zoom by the name Angel followed by a number, will assist online. If any speaker is missing or has a technical problem, the chair can pass to the next speaker and come back later to resume if possible. Detailed indications for the session chairs of both virtual and hybrid sessions can be found on the website.

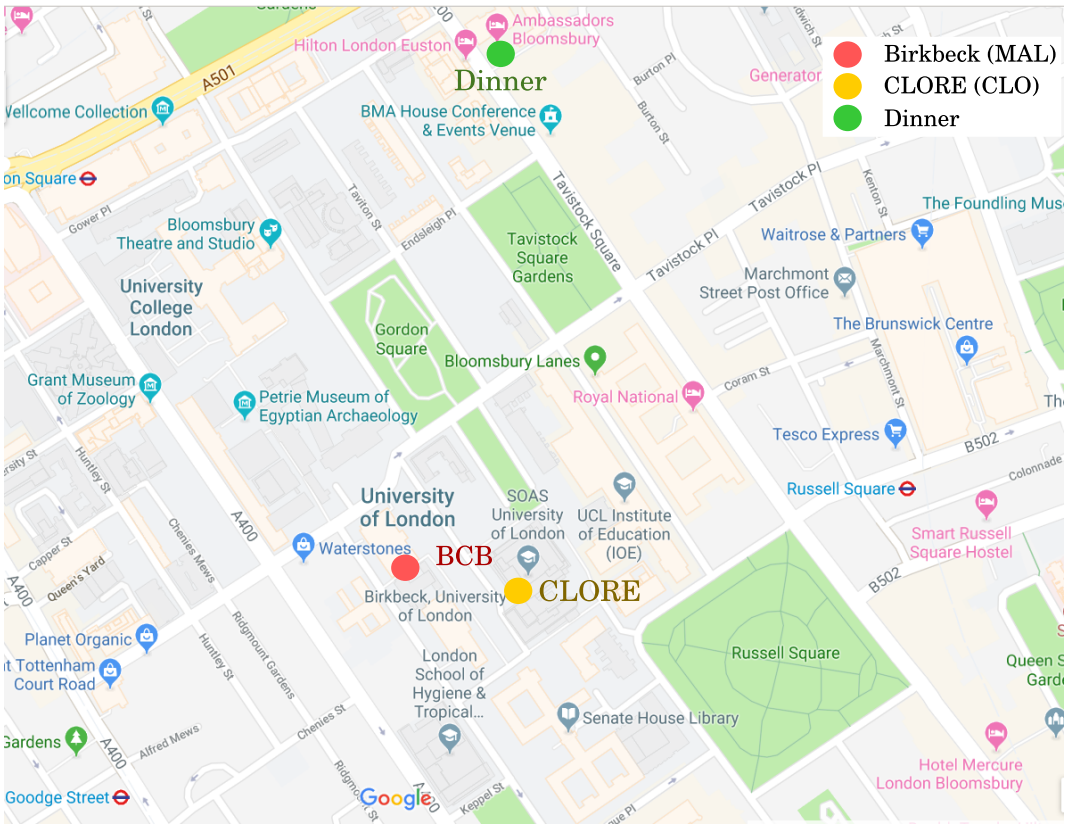
Test session

A test session will be set up for Saturday, 19th August 2023, from 15:00 to 15:30 UTC+1 (BST/UK time). The participants will be able to enter the virtual room R01 in the programme to test their presentations, video, micro and audio. Detailed indications for the test sessions can be found on the website.

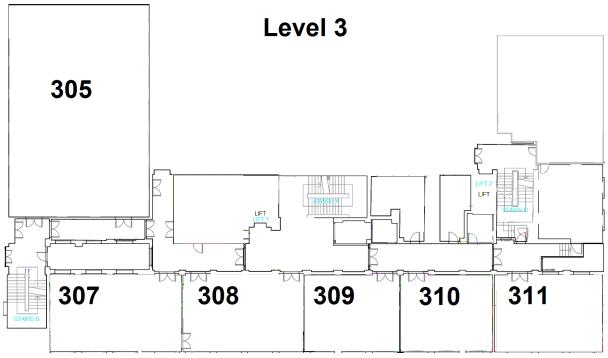
COMPSTAT venue



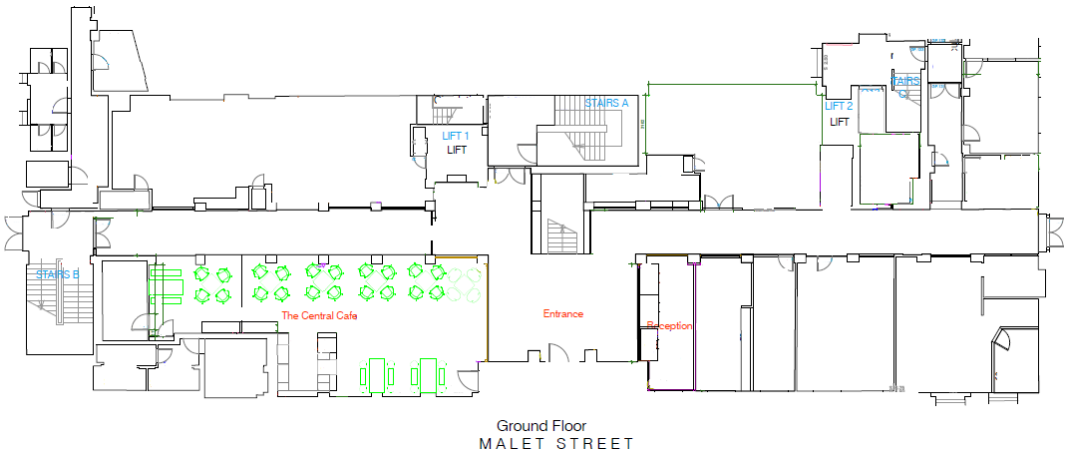
Venues and nearby area



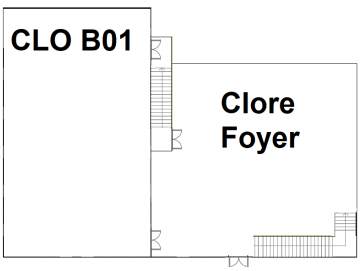
Floor 3 Birckbeck Central Building (BCB)



Ground floor Birckbeck Central Building (BCB)



Basement of Clore Management Centre



Contents

Tuesday 22.08.2023 09:00 - 10:00 Room: CLO B01 Chair: Maria Brigida Ferraro Keynote talk I

Multi-objective optimisation of restricted randomised designs

Speaker: **Kalliopi Mylona, King's College London, United Kingdom**

Scientists can now address scientific issues of increasing complexity thanks to modern experiments. Often, factors in experiments have levels that are more difficult to set than others, and in these cases, using a restricted randomised design offers a solution. There are numerous approaches to find optimal designs that focus just on maximising a specific criterion. To tackle the drawbacks of one-objective optimisation, multi-criteria techniques have been developed. We will present novel tools for the design of optimal experiments for multiple criteria, and we will show their application to motivating examples from the industry and pharmaceutical sciences.

Tuesday 22.08.2023 18:00 - 18:55 Room: CLO B01 Chair: Francisco Javier Rubio Keynote talk II

The historical role of energy in UK inflation and productivity

Speaker: **David Hendry, University of Oxford, United Kingdom**

The purpose is to model UK price and wage inflation, productivity and unemployment over a century and a half of data, selecting dynamics, relevant variables, non-linear reactions and location and trend shifts using indicator saturation estimation. The four congruent econometric equations highlight complex interacting empirical relations. The production function reveals a major role for energy inputs additional to capital and labor, and although the price inflation equation shows a small direct impact of energy prices, the substantial rise in oil and gas prices seen by mid-2022 contribute half of the increase in price inflation. We find empirical evidence for non-linear adjustments of real wages to inflation: a wage-price spiral kicks in when inflation exceeds about 6–8% p.a. We also find an additional non-linear reaction to unemployment, consistent with involuntary unemployment. A reduction in energy availability simultaneously reduces output and exacerbates inflation.

Friday 25.08.2023 12:10 - 13:10 Room: BCB 307 Chair: David Weston Keynote talk III

Subject prevalence in documents based on topic modeling

Speaker: **Ana Colubi, University of Giessen, Germany** Louisa Kontoghiorghes

A metric to quantify the relevance of specific subjects within a text is considered. The metric can be used to track the evolution of a subject in a series of documents or to measure the statistical impact of a given text in related literature. To this aim, text mining tools are combined with Bayesian and frequentist statistical methods. First, topic modeling is suggested to be employed to identify relevant topics. The derived models are used to quantify the relative importance of a subject defined through a given set of terms, or keywords, by employing Bayesian techniques. Then, bootstrap two-sample tests are proposed to compare subjects' prevalence in two groups of documents. Illustrative empirical results are provided.

Tuesday 22.08.2023 10:30 - 12:30 Parallel Session B – COMPSTAT2023

CO112 Room Virtual room R01 TUTORIAL I **Chair: Alessandra Luati**

C0179: **Dynamic models for multiple quantiles**
Presenter: **Alessandra Luati**, Imperial College London, United Kingdom
Recent developments in models for dynamic multiple quantiles are discussed. The baseline semiparametric model introduced recently, based on quantile spacings and score-type updates, is reviewed and extended to account for: heterogeneous tail behaviour, cross-tail effects, and exogenous variables. The extensions result in a flexible class of models ensuring that quantiles do not cross in finite samples and that extreme quantiles are estimated based on information coming from all the regions of the underlying conditional distribution. M-estimation is carried out, and the asymptotic properties of the estimators are discussed. Open problems and illustrations conclude the tutorial.

CO015 Room BCB 311 CATEGORICAL AND HIGH-DIMENSIONAL DATA ANALYSIS **Chair: Mark De Rooij**

C0216: **The impact of (un)congenial multiple imputation approaches on GPAbin biplots**
Presenter: **Johane Nienkemper-Swanepoel**, Stellenbosch University, South Africa
Co-authors: Niel Le Roux, Sugnet Lubbe
Multiple imputation is considered a superior technique for handling missing data. This approach results in multiple completed data sets, which are analysed separately by means of standard complete data techniques. Estimates from the separate analyses are combined using suitable combining rules, referred to as Rubin's rules. In the context of exploratory analysis, GPAbin biplots enable the unified visualisation of the individual plots constructed from multiple imputed data sets. This visualisation approach combines configurations by means of generalised orthogonal Procrustes analysis (GPA) followed by the application of Rubin's rules (-bin) on the aligned configurations. In the context of multivariate categorical data, the configurations are multiple correspondence analysis (MCA) biplots. Multiple imputation with multiple correspondence analysis (MIMCA) is therefore regarded as a suitable benchmark approach for other imputation methods, due to the congeniality between the imputation and analysis models. MIMCA is a joint modelling multiple imputation approach, since the same imputation model is used for all variables. In a previous study, the performance of the GPAbin biplots after MIMCA has been evaluated in an extensive simulation study. Now, the performance of GPAbin under joint modelling and fully conditional specification imputation approaches is compared and discussed. Through simulation, the choice of multiple imputation approach is investigated.

C0230: **A strategy for improving the speed in tensor decomposition analysis**
Presenter: **Michele Gallo**, University of Naples L'Orientale, Italy
Tensors are powerful algebraic objects for handling huge amounts of data. To extract information from data or just for compression of data, several methods can be used to process the tensors. Here, we consider the higher-order singular value decomposition also known as the Tucker tensor decomposition. Due to the huge number of elements, large collections of deterministic and randomized algorithms have been proposed, that are supposed to be faster and/or cheaper. The goal is to propose an efficient algorithm that provides the same results as HOSVD ones. The approach proposed is based on rearranging data in a higher-order tensor, and applying eigenvalues analysis on the Gram unfolding tensor. To show the accuracy and efficiency of the algorithm proposed, we perform numerical experiments on synthetic and real-world data to directly show the strengths and weaknesses of our proposal.

C0244: **Assessing dispersion in a two-way contingency table under profile transformations and reciprocal averaging**
Presenter: **Ting-Wu Wang**, University of Newcastle, Australia
Co-authors: Eric Beh, Rosaria Lombardo
Analysing the association between the categorical variables of a two-way contingency table can often be problematic due to dispersion issues that arise from the assumption that the cell frequencies are a Poisson random variable. Despite such an assumption requiring parity between the variance and expectation, the variance is typically larger than its expectation so that over-dispersion exists in the data. Applying a power transformation to the Poisson random variable has been a popular technique to overcome the presence of over-dispersion for nearly 100 years. However, detecting the presence of over-dispersion has not been examined when using reciprocal averaging; a method used to determine scores that best discriminate the categories of the contingency table while maximising the association between its variables. Nor has the power transformation of the data been examined for reciprocal averaging. Therefore, we shall be considering an index of dispersion that monitors for any dispersion issues that exist in the contingency table when applying a power transformation to the data. We shall also be assessing how variations in the power impact on the row and column scores obtained when applying a reciprocal averaging to the transformed cells of the contingency table.

C0253: **Logistic multidimensional data analysis for ordinal response variables using a cumulative link function**
Presenter: **Mark De Rooij**, Leiden University, Netherlands
A multidimensional data analysis framework is presented for the analysis of ordinal response variables. We assume a continuous latent variable underlying the ordinal variables, leading to cumulative logit models. The framework includes unsupervised methods when no predictor variables are available and supervised methods when predictor variables are available. We distinguish between dominance variables and proximity variables, where dominance variables are analyzed using inner product models, whereas the proximity variables are analyzed using distance models. An expectation-majorization-minimization algorithm is derived for the estimation of the parameters of the models. We illustrate our methodology with data from the International Social Survey Programme.

C0287: **Association-based distances for categorical and mixed-type data**
Presenter: **Alfonso Iodice D Enza**, Universita di Napoli Federico II, Italy
Co-authors: Michel van de Velden, Carlo Cavicchia, Angelos Markos
Several statistical methods are based on distances, that is, the quantification of the differences among observed values in a set of attributes. The definition of distance is not unique as it depends on the attributes describing the observations, and on the problem at hand. Distances between continuous observations result from the aggregation of attribute-wise differences, and attribute correlations may or may not be taken into account. For categorical observations, simplistic mis-matches counting aside, the definition of distance/dissimilarity is less intuitive, nor it is unique: several distance measures have been proposed, and choosing one is subjective, more so in unsupervised learning. In the mixed-data case, distances computation requires further choices, mostly to balance out the impact of the continuous and categorical attributes. An association-based distance for categorical and mixed data is proposed that takes into account the categorical/categorical and continuous/categorical attributes relations.

CC073 Room BCB 307 SPATIAL STATISTICS **Chair: Klaus Nordhausen**

C0152: **Spatial smoothing using graph Laplacian penalized filter**
Presenter: **Hiroshi Yamada**, Hiroshima University, Japan
A filter is considered for smoothing spatial data. Since smoothing coincides with detrending, spatial detrending is also considered. The filter consists of a quantity analogous to Geary's c, which is one of the most prominent measures of spatial autocorrelation. In addition, the quantity can be represented by a matrix called the graph Laplacian in the spectral graph theory/linear algebraic graph theory. We show mathematically how spatial data become smoother as a parameter called the smoothing parameter increases from 0 and fully smoothed as the parameter goes to infinity,

except for the case where spatial data are originally fully smoothed. We also illustrate the results numerically. In the numerical illustration, we demonstrate how the generalized cross-validation criterion works for specifying the smoothing parameter. Finally, as supplementary investigations, we examine how the sum of squared residuals and effective degrees of freedom vary with the smoothing parameter.

C0316: **Analyzing regional suicide patterns in Japan before and after the COVID-19 pandemic and usage of generative AI for EBPM**
Presenter: **Takafumi Kubota**, Tama University, Japan

The purpose is to explore regional trends in suicide deaths in Tokyo, Japan, before and after the COVID-19 pandemic. In addition, another objective is to generate evidence for policy-making by each municipality using generative AI and to examine the accuracy of such evidence. The data covered in this study are data on suicide by a municipality in Tokyo in 2016 and 2021. They are based on the "Basic Data on Suicide in Local Communities" of the statistics on suicide compiled by the Ministry of Health, Labour and Welfare. In addition to the suicide death rate, suicide items include age, occupation, location, etc., with 71 variables, including region names. The methodology first compares suicide data for 2016 and 2021 by visualization. Next, regions with exceptionally high rates of each item are identified. The method used to identify regions is spatial clustering of geographically referenced attributes. Moreover, the focus is placed on a single city. Here, Fuchu City in Tokyo is used as the object of comparison. The comparison targets are the annual changes in Fuchu City in 2016 and 2021 (A) and Tokyo and Fuchu City in 2021 (B). The author submits the data of 71 variables for A and B to ChatGPT (version 4), one of the generative AIs, with instructions to detect differences and have it generate a document about the results. The content of the generated documents is verified to be accurate and to provide evidence that can be used for policy-making.

C0382: **Commuting and the spread of infectious diseases: A spatio-temporal analysis of influenza in Germany**
Presenter: **Manuel Stapper**, WWU Muenster, Germany

The spread of infectious diseases is a major public health concern, with population mobility and commuting as two important driving factors. Frequently used epidemiologic models are extended by incorporating commuting network data to examine the spread of Influenza in Germany between 2001 and 2019. It is demonstrated that the established finding of human mobility to follow a power-law is reflected in a power-law of contact probabilities under certain assumptions. A novel type of weight matrix for the spatio-temporal model is built from contact probabilities, capable of reflecting the number of potentially hazardous interactions between residents of two administrative districts. The results enable an exploration of the spread's typical paths and the assessment of policy measures through forecasting under different conditions.

C0383: **Flexible basis representations for modeling high-dimensional hierarchical spatial data**
Presenter: **Seiyon Lee**, George Mason University, United States
Co-authors: Remy MacDonald

Nonstationary and non-Gaussian spatial data are prevalent across many fields (e.g., counts of animal species, disease incidences in susceptible regions, and remotely-sensed satellite imagery). Due to modern data collection methods, the size of these datasets has grown considerably. Spatial generalized linear mixed models (SGLMMs) are a flexible class of models used to model nonstationary and non-Gaussian datasets. Despite their utility, SGLMMs can be computationally prohibitive for even moderately large datasets. To circumvent this issue, past studies have embedded the nested radial basis function into the SGLMM. However, two crucial specifications (knot locations and bandwidths), which directly affect model performance, are generally fixed prior to model fitting. We propose a novel algorithm to model large nonstationary and non-Gaussian spatial datasets using adaptive radial basis functions. Our approach: (1) partitions the spatial domain into subregions; (2) selects a carefully curated set of basis knot locations within each partition; and (3) models the latent spatial surface using partition-varying and data-driven (adaptive) basis functions. Through an extensive simulation study, we show that our approach provides more accurate predictions than a competing method while preserving computational efficiency. We also demonstrate our approach on two environmental datasets that feature incidences of parasitic plant species and counts of bird species in the United States.

C0384: **Inference for spatial autoregressive models using stochastic gradient descent**
Presenter: **Ji Meng Loh**, New Jersey Institute of Technology, United States
Co-authors: Gan Luan

The stochastic gradient descent (SGD) procedure is considered to fit spatial auto-regressive models to lattice data, incorporating a recently developed perturbation method to obtain standard errors in addition to model parameter estimates. We derive the SGD update equations and will present the results of a simulation study to examine the empirical coverage of confidence intervals constructed using the perturbation procedure.

CC062 Room BCB 308 BIOSTATISTICS Chair: Stefan Van Aelst

C0326: **A multi-state modeling with Poisson regression utilizing grouped data in a radiation epidemiological study**
Presenter: **Munechika Misumi**, Radiation Effects Research Foundation, Japan
Co-authors: Hiromi Sugiyama

By treating the event occurrence hazard as a piecewise constant function, we can perform survival time analysis using Poisson regression. This approach enables flexible modeling of the baseline hazard and accommodates multiple time-dependent covariates within the generalized linear/nonlinear model framework. In radiation epidemiology, particularly for the analysis of long-term follow-up data from epidemiological cohorts, Poisson regression has become a standard methodology. To implement this regression, we must stratify the time-to-event dataset to create grouped data, a complexity that has hindered the execution of more flexible survival analyses in the field. Furthermore, traditional Poisson regression assumes independent and non-informative censoring, an assumption that is invalidated when subjects experiencing a competing event are censored. In response to these challenges, we propose the use of a multi-state model based on Poisson regression with grouped data to analyze long-term follow-up data. As a motivating example, we use data from the Life Span Study (LSS) of Japanese atomic bomb survivors. The LSS dataset comprises over 120,000 survivors from Hiroshima and Nagasaki, with a follow-up length exceeding 50 years. The analysis provided a comprehensive interpretation of the relationship between radiation exposure and incidence of rectal cancer and adenoma, which are multiple events in a carcinogenesis pathway.

C0379: **Generation of synthetic mixed data for multiple sclerosis patients: Application to gait data and EDSS score**
Presenter: **Klervi Le Gall**, Nantes University, France
Co-authors: Lise Bellanger, Aymeric Stamm, David Laplaud

Gait analysis is a key factor in the understanding and care of multiple sclerosis. To analyse gait, we developed a biomarker which characterises the rotation of the hip of an individual during an average gait cycle using unit quaternion time series (QTS) collected with a motion sensor. To complete the data, EDSS scores (qualitative markers of disease progression) were assessed by neurologists. We developed a promising clustering method to group patients with similar gait impairments using a small database. We need a larger volume of data to assess its robustness and validate the method. We propose a sound statistical method to generate synthetic mixed data, including QTS and EDSS scores which best resembles the original data set while keeping anonymity and preserving inertia. To the best of our knowledge, this is the first time that such a method is proposed for functional data evaluation in the Lie group of 3D rotations. The proposed approach builds a synthetic data set by mixing multivariate functional PCA, multivariate analysis of mixed data and nearest neighbours weighting. We will show the relevance of our approach using a sample of multiple sclerosis patients from a clinical study conducted in collaboration with the University Hospital of Nantes.

C0340: **Dimensionality reduction for multi-omics data using the Freeman-Tukey transformation**
Presenter: **Hiroshi Kobayashi**, Doshisha University, Japan
Co-authors: Masaaki Okabe, Hiroshi Yadohisa

The data integration method is useful for understanding the complex patterns between multiple datasets. When multiple datasets contain complementary information, data integration can remove noise and discover common structures. The Co-Inertia Analysis (CIA) method describes the relationships between multiple datasets by maximizing the covariance between them. Therefore, it captures the structure associated with multiple datasets and enables low-dimensional visualization without losing the unique information contained in each dataset. When applied to overdispersed data or highly sparse count data, CIA often requires a combination with logarithmic transformation. However, dimension reduction through logarithmic transformation may distort the structure of the data. We propose CIA, using the Freeman-Tukey transformation. This technique is able to better capture low-dimensional structures that preserve the structure of the original count data. It also enables dimension reduction of multiple datasets without compromising the biological structure, while integrating real multi-omics data. Moreover, it allows for downstream analysis that accurately captures the inherent biological features, resulting in deeper insights into complex biological characteristics.

C0319: **Estimating the linear relation between variables that are never jointly observed: An application to in vivo experiments**
Presenter: **Polina Arsenteva**, Institut de Mathematiques de Bourgogne, France
Co-authors: Mohamed Amine Benadjaoud, Herve Cardot

The motivation comes from in vivo experiments in which different measurements are performed on different animals. Thus, the variables of interest can never be observed simultaneously, making the task of estimating the linear regression coefficients challenging. Assuming that the global experiment can be decomposed into subpopulations (corresponding, for example, to different doses of a treatment substance) with distinct first moments, we propose different estimators of the linear regression, which take into account this additional information. We consider a method of moments approach as well as an approach based on optimal transport theory. These estimators are proved to be consistent as well as asymptotically Gaussian under weak hypotheses. Bootstrap techniques are shown to give consistent confidence intervals for the estimated parameter. A Monte Carlo study is conducted to assess and compare the finite sample performances. Finally, the proposed approaches are illustrated in the context of radiobiology, namely a preclinical study on mice investigating the multiscale correlation between the inflammation process (gene expression) and lung injury (septal thickening) appearing after irradiation, with a further comparison of the results across several irradiation configurations.

C0325: **Network inference and robust clustering on high-dimensional data to investigate molecular heterogeneity in glioma**
Presenter: **Roberta Coletti**, Center for Mathematics and Applications (NOVA Math), NOVA School of Science and Technology, Portugal
Co-authors: Marta Lopes, Sofia Martins

Gliomas are a family of brain tumors that generally exhibit a low patient survival rate. Finding novel targets for personalized therapies necessitates a deeper knowledge of molecular underpinnings in various glioma types, which can be achieved by statistical and machine learning methods applied to the high-dimensional data sets nowadays generated. We propose a mathematical workflow to investigate differences and similarities in the three main glioma types: astrocytoma, oligodendroglioma, and glioblastoma. Based on gene expression data from The Cancer Genome Atlas, updated following the 2016 glioma classification guidelines, we estimated a sparse gene network for each glioma type by applying the joint graphical lasso algorithm. This led to a network-based variable selection, which was validated through robust sparse K-means clustering in different cases of study, to detect relevant genes in the separation of samples between classes in an unsupervised way. The outcomes disclose molecular differences between glioblastoma and the other glioma subtypes and point to potential novel glioma biomarkers needing further biological validation. The sets of selected variables appeared meaningful in the identification of robust clusters, though not totally in agreement with the preassigned diagnostic labels. This result supports further efforts towards the revision of the criteria for glioma classification.

CC030 Room BCB 310 TIME SERIES Chair: Matus Maciak

C0185: **Linear law-based feature space transformation**
Presenter: **Marcell Tamas Kurbucz**, Wigner Research Centre for Physics | Corvinus University of Budapest, Hungary
Co-authors: Antal Jakovac, Peter Posfay

The aim is to facilitate uni- and multivariate time series classification tasks using linear law-based feature space transformation (LLT). This new algorithm first splits the instances into training and test sets. Then, it identifies the governing patterns (laws) of each input sequence in the training set by applying time-delay embedding and spectral decomposition. Finally, it uses the laws of the training set to transform the feature space of the test set. These calculation steps have a low computational cost and the potential to form a learning algorithm. The application of LLT is illustrated using datasets from the fields of human activity recognition, price movement prediction, and electrocardiogram signal classification.

C0186: **Regularized nonlinear regression with dependent errors and its application to a biomechanical model**
Presenter: **Wei-Ying Wu**, National Dong Hwa University, Taiwan

A biomechanical model often requires parameter estimation and selection in a known but complicated nonlinear function. Motivated by observing that the data from a head-neck position tracking system, one of the biomechanical models, show multiplicative time-dependent errors, we develop a modified penalized weighted least squares estimator that can handle such error structure. The proposed method can also be applied to a model with non-zero meantime-dependent additive errors. Asymptotic properties of the proposed estimator are investigated. A simulation study demonstrates that the proposed estimation performs well in both parameter estimation and selection with time-dependent error. The analysis and comparison with an existing method for head-neck position tracking data show better performance of the proposed method in terms of the variance accounted for (VAF).

C0364: **High-dimensional high-frequency time series prediction model solved with a mixed integer optimisation method**
Presenter: **Nazgul Zakiyeva**, Technische Universitat Berlin, Germany

A network functional autoregressive model is studied for large-scale network time series. We approach the estimation of the proposed model using a mixed integer optimisation method. By including the high-dimensional curves, the proposed model captures both serial and cross-sectional dependence in the functional time series network. We illustrate our methodology on large-scale natural gas network data where our model provides more accurate several days-ahead hourly out-of-sample forecasts of the gas in- and out-flows compared to alternative prediction models.

C0386: **Fused lasso nearly-isotonic signal approximation in general dimensions**
Presenter: **Vladimir Pastukhov**, Chalmers University of Technology, Sweden

The purpose is to introduce and study fused lasso nearly-isotonic signal approximation, which is a combination of fused lasso and generalized nearly-isotonic regression. We show how these three estimators relate to each other and derive solutions to the general problem. Our estimator is computationally feasible and provides a trade-off between monotonicity, block sparsity and goodness-of-fit. Next, we prove that fusion and near-isotonisation in the one-dimensional case can be applied interchangeably. Also, we derive an unbiased estimator of the degrees of freedom of the estimator.

C0389: **The modified conditional sum-of-squares estimator for fractionally integrated models**
Presenter: **Mustafa Kilinc**, WHU - Otto Beisheim School of Management, Germany
Co-authors: Michael Massmann

The aim is to analyze the influence of estimating a constant term on the bias of the conditional sum-of-squares (CSS) estimator of the fractional parameter, say d, and other parameter estimates of the short-run dynamics in a stationary or non-stationary time series model. We first consider a "type II" ARFIMA(0,d,0) model including a constant term and derive an expression for the leading bias term of $\hat{d}$. We show that we can easily remove the bias in $\hat{d}$ that occurs due to the presence of a constant term by a simple modification of the CSS objective function. Consequently, the estimated fractional parameter, say $\hat{d}_m$, behaves on average the same as if we had known the true value of the constant term, discounting the higher-order bias terms. We call this new estimator the modified conditional sum-of-squares estimator (MCSS). The remaining part of the leading bias of $\hat{d}_m$ is pivotal and can be completely eliminated by a simple bias correction. We later generalize our analysis to the case where the short-run dynamics take a more general structure than the simple $i.i.d.$ shocks. The importance of bias correction is highlighted through empirical illustrations in macroeconomics and hydrology time series.

Tuesday 22.08.2023 14:15 - 15:45 Parallel Session C – COMPSTAT2023

CO017 Room Virtual room R01 STATISTICAL METHODS FOR SPATIAL AND SPATIO-TEMPORAL DATA Chair: Hsin-Cheng Huang

C0165: **Scalable semiparametric spatio-temporal regression for large data analysis**
Presenter: **Jun Zhu**, University of Wisconsin - Madison, United States

With the rapid advances in data acquisition techniques, spatio-temporal data are becoming increasingly abundant in a diverse array of disciplines. Spatio-temporal regression methodology is developed for analyzing large amounts of spatially referenced data collected over time, motivated by environmental studies utilizing remotely sensed satellite data. In particular, a semiparametric autoregressive model is specified without the usual Gaussian assumption and a computationally scalable procedure is devised that enables the regression analysis of large datasets. The model parameters are estimated by maximum pseudo-likelihood, and the computational complexity can be reduced from cubic to linear of the sample size. Asymptotic properties under suitable regularity conditions are further established that inform the computational procedure to be efficient and scalable. A simulation study is conducted to evaluate the finite-sample properties of the parameter estimation and statistical inference. The methodology is illustrated by a dataset with over 2.96 million observations of annual land surface temperature, and a comparison with an existing state-of-the-art approach to spatio-temporal regression highlights the advantages of our method.

C0215: **On minimum contrast method for multivariate spatial point processes**
Presenter: **Junho Yang**, Academia Sinica, Taiwan

Compared to widely used likelihood-based approaches, the minimum contrast (MC) method is a computationally efficient method for the estimation and inference of parametric stationary point processes. This advantage becomes more pronounced when analyzing complex point process models, such as multivariate log-Gaussian Cox processes (LGCP). Despite its practical advantages, there is very little work on the MC method for multivariate point processes. The aim is to introduce a new MC method for parametric multivariate stationary spatial point processes. A contrast function is calculated based on the trace of the power of the difference between the conjectured K-function matrix and its nonparametric unbiased edge-corrected estimator. Under standard assumptions, the asymptotic normality of the MC estimator of the model parameters is derived. The performance of the proposed method is illustrated with bivariate LGCP simulations and real data analysis of a bivariate point pattern of the 2014 terrorist attacks in Nigeria.

C0359: **ZIP-like models for spatial count processes**
Presenter: **Chun-Shu Chen**, National Central University, Taiwan
Co-authors: Chung-Wei Shen

Spatial count responses with an excessive number of zeros and a set of covariates are common. To alleviate deviations from model assumptions, we propose a spatial zero-inflated Poisson-like methodology to model this type of data which only relies on the first two moments of responses. We design an iterative estimation procedure to estimate regression coefficients and variograms under the generalized estimating equation framework. Also, the stabilization of estimators is evaluated via a block jackknife method. After parameter estimation, a criterion based on the mean squared error of the estimated mean structure is proposed to select covariates. Numerical results show the effectiveness of the proposed methodology.

C0269: **Nonstationary spatial modeling, estimation, and prediction using a divide-and-conquer approach**
Presenter: **Hsin-Cheng Huang**, Academia Sinica, Taiwan

Spatial data over a large domain generally shows nonstationary spatial covariance characteristics. However, estimating a nonstationary covariance function from a single realization of data is challenging, and the computation of the optimal spatial prediction is intractable when the dataset is massive. We initially propose a method for visualizing nonstationary covariance structures and introduce a statistical test for spatial stationarity. Upon detection of nonstationarity, we propose a segmentation technique that decomposes the spatial domain into K subregions wherein the process is approximately stationary. Additionally, we consider a stationary process for each of these K subregions and subsequently develop a novel nonstationary model that employs a linear combination of these processes with spatially varying weights. Contrary to independent stationary models, our approach treats the K stationary processes as interdependent and represents them using a multivariate Matern covariance model. The proposed nonstationary model showcases flexibility, morphing into a globally stationary process when all stationary components exhibit a shared spatial covariance structure. Finally, we propose a divide-and-conquer strategy for fast spatial prediction. The effectiveness of our approach is demonstrated through numerical experiments.

CO016 Room BCB 307 BAYESIAN METHODS Chair: Toshiaki Watanabe

C0275: **Time-varying parameter heterogeneous autoregressive model with stochastic volatility**
Presenter: **Toshiaki Watanabe**, Hitotsubashi University, Japan
Co-authors: Jouchi Nakajima

The heterogeneous autoregressive (HAR) model is known to perform well in volatility forecasting. This model formulates realized volatility (RV) as a function of past RVs with different frequencies such as daily, weekly and monthly RVs. The HAR model is extended such that the coefficients of daily, weekly and monthly RVs and the error variance may change over time. The coefficients and the log of the error variance are assumed to follow first-order autoregressive processes. A Bayesian method using an efficient Markov chain Monte Carlo is developed for the analysis of the proposed model. An empirical application with the RV calculated using the 5-minute returns of the Nikkei 225 stock index is provided.

C0290: **Bayesian model selection among dispersed integer-valued time series models**
Presenter: **Feng Chi Liu**, Feng Chia University, Taiwan
Co-authors: Cathy W-S Chen, Hsiao-Han Hsu

The focus is on a class of integer-valued time series models with over-dispersion and extends those models to generalized forms. These new models include: (1) dispersed INGARCH models incorporating negative binomial, double Poisson, or generalized Poisson, and (2) double log-form INGARCH model. The latter model avoids over-restrictions in the parameter space. We perform parameter estimations and model selection within the Bayesian framework, employ adaptive Markov chain Monte Carlo (MCMC) sampling schemes, and calculate the deviance information criterion (DIC) for model selection. Simulation studies demonstrate that the proposed method accurately estimates the model parameters with reliable MCMC samples. Taking monthly crime counts in Bankstown, New South Wales, Australia, for an empirical illustration, the findings show the ability to select the promising models among the competing models in terms of DIC.

C0291: **A complete Bayesian degradation analysis based on inverse Gaussian processes**
Presenter: **Tsai-Hung Fan**, National Central University, Taiwan
Co-authors: Yi-Shain Dong, Chien-Yu Peng

Degradation models are constructed for the observations of a quality characteristic related to the failure time of products. The failure time inference of the product is derived based on the first passage time to a specific threshold for the selected degradation model. The Bayesian analysis incorporated with valuable prior information from expert opinion or experience is a useful approach, in particular for small sample sizes. However, most Bayesian research focuses more on the degradation model than the failure time inference. Bayesian predictive analysis is used based on the inverse Gaussian process with conjugate priors to deduce the failure time inference. The posterior inference of the parameters for the fixed-effect linear

degradation model is derived in closed forms, and the full conditional posteriors are developed for the random-effect models using hierarchical modeling. The failure time inference associated with the degradation model and its goodness-of-fit test is suggested from a complete Bayesian perspective. The proposed failure time inference can be used for other degradation models with random-effect. An illustrative example demonstrates the feasibility and advantages of the proposed Bayesian approach.

C0315: **Rating of players by Laplace approximation and dynamic modeling**
Presenter: **Chiu-Hsing Weng**, National Chengchi University, Taiwan

The Elo rating system is a simple and widely used method for calculating players' skills from paired comparison data. Many have extended it in various ways. Yet the question of updating players' variances remains to be further explored. We address the issue of variance update by using the Laplace approximation for posterior distribution, together with a random walk model for the dynamics of players' strengths. The random walk model is motivated by the Glicko system, but here we assume nonidentically distributed increments to take care of player heterogeneity. Experiments on men's professional matches showed that the prediction accuracy slightly improves when the variance update is performed. It also showed that young players' strengths may be better captured with the variance update.

CO028 Room BCB 310 RANK-BASED INFERENCE, FEATURE SELECTION, AND DATA CONSOLIDATION Chair: Michael Georg Schimek

C0195: **Modeling of preference data with multiple network views**
Presenter: **Philip Yu**, The Education University of Hong Kong, Hong Kong
Co-authors: Yipeng Zhuang

Nowadays, people are connected to many networks. Their preferences for items (such as ratings and rankings) might be affected by some of these networks. We will introduce some methods for modeling preference data with multiple network views. We will apply our proposed models to some real-world movie recommendation datasets. Empirical results demonstrate that our models achieve significant improvements in preference prediction over other existing models.

C0256: **Rank-based Bayesian joint variable selection and clustering of genome-wide transcriptomic data**
Presenter: **Valeria Vitelli**, University of Oslo, Norway

The use of ranks in genomics is naturally linked to the underlying biological question, since one is often interested in overly-expressed genes in a given pathology. When aiming at analysing transcriptomic patient data for cancer subtype discovery, we have already successfully proposed to use a mixture-based clustering approach rooted in Bayesian Mallows models (BMM). BMM is able to handle heterogeneous patient data, and to both produce estimates of the consensus ranking of the genes shared among samples in the same cluster, and to fill in missing data via data augmentation. However, BMM is computationally intensive, thus relying on pre-selecting around 1000 genes to be used in the analysis. A lower-dimensional version of BMM (lowBMM) that scales to genome-wide transcriptomic data has also been proposed and used in the context of cancer genomics; however, lowBMM does not perform clustering. We now propose to perform genome-wide cancer subtyping of transcriptomic patient data via a Bayesian mixture of Mallows models that combines BMM and lowBMM. The model jointly performs clustering and variable selection, thus selecting the genes best representing the structural patterns of expression characterising each subtype. We study the performance of the method via simulations, and show the results of a pan-cancer analysis.

C0255: **Improving the stability of tree-based feature importance ranks via consensus signals**
Presenter: **Bastian Pfeifer**, Medical University of Graz, Austria
Co-authors: Michele La Rocca, Michael Georg Schimek

Feature (variable) selection methods are used to detect the most important features (variables) within the data. Tree-based models such as Random Forests are particularly suited for such purposes as they include a model in-build mechanism to quantify feature importance. For this reason, specialized feature selection techniques are often based on tree-based models. The underlying stochastic sampling strategies of these approaches, however, are causing unstable feature importance ranks, especially when the number of trees is low. We investigate to which extent the observed stochastic fluctuation can be consolidated via consensus ranks. We propose to compute consensus values from multiple feature selection runs to stabilise the feature ranks, where each run forms a ranked list. We compare standard rank aggregation techniques with more sophisticated approaches, where consensus ranks are inferred via signal reconstruction using convex optimization. The latter is especially suited not only to stabilize the obtained feature importance ranks, but also to verify the most important subset of features.

C0237: **Bootstrap inference for signal reconstruction from multiple ranked lists**
Presenter: **Michele La Rocca**, University of Salerno, Italy
Co-authors: Bastian Pfeifer, Michael Georg Schimek

Statistical ranking procedures are widely used to rate objects' relative quality or relevance across multiple assessments. Beyond rank aggregation, estimating the usually unobservable latent signals that inform a consensus ranking is interesting. Under the only assumption of independent assessments, we have introduced an indirect inference approach via convex optimisation, which is computationally efficient even when the number of assessors is much lower than the number of objects to rank. Notably, the novel signal estimator can be written as a weighted estimator, which opens the possibility of using weighted bootstrap schemes to implement efficient resampling procedures. Those procedures are key in gaining inference on the unknown signals (by computing standard errors or confidence intervals) or testing signal differences between groups. Within this framework, we compare alternative weighted bootstrap schemes (namely, Poisson, Multinomial and Dirichlet) for their different computational burden and ability to approximate the unknown sampling distribution of the signal estimators accurately. The bootstrap procedures will be evaluated, by Monte Carlo simulation, on different scenarios with increasing problem complexity, including several combinations of the number of assessors and the number of objects to rank.

CO026 Room BCB 311 CMSTATISTICS SESSION: STATISTICAL ANALYSIS OF COMPLEX DATA Chair: Enea Bongiorno

C0155: **Kernel ordinary differential equations**
Presenter: **Lexin Li**, University of California Berkeley, United States

Ordinary differential equation (ODE) is widely used in modeling biological and physical processes in science. We propose a new reproducing kernel-based approach for estimation and inference of ODE given noisy observations. We do not assume the functional forms in ODE to be known, or restrict them to be linear or additive, and we allow pairwise interactions. We perform sparse estimation to select individual functionals, and construct confidence intervals for the estimated signal trajectories. We establish the estimation optimality and selection consistency of kernel ODE under both the low-dimensional and high-dimensional settings, where the number of unknown functionals can be smaller or larger than the sample size. Our proposal builds upon the smoothing spline analysis of variance (SS-ANOVA) framework, but tackles several important problems that are not yet fully addressed, and thus extends the scope of existing SS-ANOVA as well. We demonstrate the efficacy of our method through numerous ODE examples.

C0161: **Spatio-temporal deepkriging for interpolation and probabilistic forecasting**
Presenter: **Pratik Nag**, King Abdullah University of Science and Technology, Saudi Arabia

Gaussian processes (GP) and Kriging are widely used in traditional spatio-temporal modelling and prediction. These techniques typically presuppose that the model has a parametric covariance structure and the data are observed from a stationary GP. However, processes in real-world applica-

tions often exhibit non-Gaussianity and nonstationarity with complex dependence structures. Moreover, it is well-known that the likelihood-based inference for GPs is computationally expensive and thus prohibitive for large datasets. We propose a deep neural network (DNN) based two-stage model for spatio-temporal interpolation and forecasting. Interpolation is performed in the first step, which utilizes a dependent DNN with the embedding layer constructed by spatio-temporal basis functions. For the second stage, we propose to use Long-Short Term Memory (LSTM) and convolutional LSTM to forecast future observations at a given location. We adopt the quantile-based loss function in the DNN to provide probabilistic forecasting. Compared to Kriging, the proposed method does not require specifying covariance functions or making stationarity assumption, and is computationally efficient. Therefore, it is suitable for large-scale prediction of complex spatio-temporal processes. We apply our method to daily evapotranspiration data at more than 1 million locations from January 2019 to December 2021 for fast imputation of missing values and provide forecasts with uncertainties.

C0227: **A flexible bias correction method based on inconsistent estimators**

Presenter: **Stephane Guerrier**, University of Geneva, Switzerland

Co-authors: Mucyo Karemera, Samuel Orso, Maria-Pia Victoria-Feser, Yuming Zhang, Yanyuan Ma

An important challenge in statistical analysis lies in controlling the estimation bias when handling the ever-increasing data size and model complexity. For example, approximate methods are increasingly used to address the analytical and/or computational challenges when implementing standard estimators, but they often lead to inconsistent estimators. So consistent estimators can be difficult to obtain, especially for complex models and/or in settings where the number of parameters diverges with the sample size. We propose a general simulation-based estimation framework that allows the construction of consistent and bias-corrected estimators for increasing dimension parameters. The key advantage of the proposed framework is that it only requires computing a simple inconsistent estimator multiple times. The resulting Just Identified iNdirect Inference estimator (JINI) enjoys nice properties, including consistency, asymptotic normality, and finite sample bias correction, better than alternative methods. We further provide a simple algorithm to construct the JINI in a computationally efficient manner. Therefore, the JINI is especially useful in settings where standard methods may be challenging to apply, for example, in the presence of misclassification and rounding. We consider comprehensive simulation studies and analyze an alcohol consumption data example to illustrate the excellent performance and usefulness of the method.

C0239: **Exact and approximate moment derivation for probabilistic loops with non-polynomial assignments**

Presenter: **Efstathia Bura**, Vienna University of Technology, Austria

Co-authors: Andrey Kofnov, Ezio Bartocci, Marcel Moosbrugger, Miroslav Stankovic

Probabilistic programs (PPs) are modern tools to automate statistical modeling. They are becoming ubiquitous in AI applications, security/privacy protocols and stochastic dynamical system modeling. Many stochastic continuous-state dynamical systems can be modeled as probabilistic programs with nonlinear non-polynomial updates in non-nested loops. We present two methods, one approximate and one exact, to compute automatically and without sampling moment-based invariants for such probabilistic programs as a closed-form solution in loop iteration. The exact method applies to probabilistic programs with trigonometric and exponential updates and is embedded in the Polar tool. The approximate moment propagation method applies to any nonlinear random function as it exploits the theory of polynomial chaos expansion to approximate non-polynomial updates as the sum of orthogonal polynomials. This translates the dynamical system to a non-nested loop with polynomial updates, and thus renders it conformable with the Polar tool that computes the moments of all orders of the state variables. We evaluate our methods on an extensive number of examples ranging from modeling monetary policy to several physical motion systems in uncertain environments. The experimental results demonstrate the advantages of our approach with respect to the current state-of-the-art.

CC053 Room BCB 308 STATISTICAL MODELLING AND INFERENCE Chair: Mark De Rooij

C0158: **A new proposal to mitigate multicollinearity in linear regression models**

Presenter: **Aslam Muhammad**, Bahauddin Zakariya University, Pakistan

Co-authors: Shakeel Ahmad

A new general class of biased estimators is proposed which includes some popular estimators as special cases and discusses its properties for multiple linear regression models with the issue of multicollinearity. This proposal extends the existing Liu and Liu-type estimators. The performance of the proposed estimator is compared with many of the leading estimators, using the mean squared error matrix criterion. An extensive simulation study has shown an attractive performance of the proposed estimator. Finally, the application of the proposed estimator has been demonstrated with a chemical dataset.

C0173: **Unbiased estimators of the cumulants under bi-additive models**

Presenter: **Sandra Ferreira**, University of Beira Interior, Covilha, Portugal

Co-authors: Dario Ferreira

The focus is on the mixed-effects linear model, a topic that has been extensively researched in both theory and practice, particularly when the normal distribution is considered. We now consider the case where the random component of the model is assumed to follow a Gamma distribution. We focus on the challenges of using the fourth cumulant, including the need for more accurate and reliable estimates. We then discuss the estimators of the cumulants, which involve linear combinations of first-order cumulants with randomly chosen coefficients. To conclude, we present a simulation study and analyze its results in detail.

C0301: **Analysis of parameter and partial parameter impacts**

Presenter: **Serhat Guenay**, Competence Center for Clinical Trials Bremen, Germany

Building on work on counterfactual distributions, we consider the problem of defining and estimating the influence of small changes in a given statistical parameter (called the independent parameter) on another statistical parameter (called the target parameter) under the assumption that a set of statistical control parameters remains constant. Small changes in statistical parameters are realised by small changes in the distribution, using the theory of influence functions. Examples of the parameters are expected values, quantiles and regression coefficients. In the absence of control parameters, the influence of the independent parameter on the target parameter (called parameter impact) is defined by the ratio between the change in the target parameter and the change in the independent parameter when the distribution is perturbed along the independent parameter. We show that with a set of control parameters, orthogonalising the influence functions allows these parameters to be kept constant. This leads to a quantity we call the partial parameter impact. Point estimation and resampling-based statistical inference for influence and partial parameter impact are discussed. The method is illustrated with an observational study aimed at investigating factors that may affect the quality of inpatient geriatric care.

C0328: **Parameter estimation with increased precision for elliptic and hypo-elliptic diffusions**

Presenter: **Yuga Iguchi**, University College London, United Kingdom

Co-authors: Alexandros Beskos, Matthew Graham

Parameter estimation is considered for elliptic and hypoelliptic diffusions. Established approaches for likelihood-based estimation invoke a numerical time discretisation scheme for the approximation of the generally intractable transition dynamics of the Stochastic Differential Equation (SDE) over finite time periods. First, we propose two weak second-order sampling schemes to cover both the hypoelliptic and elliptic classes and generate novel transition density schemes of the SDE, i.e., approximations of the SDE transition density. We then provide a collection of analytical

and numerical results that solidifies the proposed schemes and showcases advantages from their incorporation within SDE calibration methods, under both high and low-frequency observations regime. Typically, for hypoelliptic diffusions, the proposed contrast estimator constructed from the transition density scheme achieves asymptotic normality under the weakest requirement for the step size of observations in the literature.

Tuesday 22.08.2023 16:15 - 17:45 Parallel Session D – COMPSTAT2023

CI003 Room BCB 307 BAYESIAN NONPARAMETRIC METHODS AND COMPUTING Chair: Michele Guindani

C0400: **Bayesian nonparametric inference for conditional vine copulas**
Presenter: **Luciana Dalla Valle**, University of Plymouth, United Kingdom
Co-authors: Rosario Barone

In recent years, conditional copulas, that allow dependence between variables to vary according to the values of one or more covariates, have attracted increasing attention. However, the literature mainly focused on the bivariate case, since the constraints on the multivariate copulas correlation matrices would make the specifications of covariates arduous. In high dimension, vine copulas offer greater flexibility compared to multivariate copulas, since they are constructed using bivariate copulas as building blocks. We present a novel inferential approach for multivariate distributions, which combines the flexibility of vine constructions with the advantages of Bayesian nonparametrics, not requiring the specification of parametric families for each pair copula. Expressing multivariate copulas using vines allows us to easily account for covariate specifications driving the dependence between response variables. We specify the vine copula density as an infinite mixture of Gaussian copulas, defining a Dirichlet process prior on the mixing measure, and performing posterior inference via Markov chain Monte Carlo sampling. Our approach is successful for clustering as well as for density estimation. We carry out simulation studies and apply the proposed approach to analyse a veterinary dataset and investigate the impact of natural disasters on financial development.

C0401: **Mixture modeling via vectors of normalized independent finite point processes**
Presenter: **Federico Camerlenghi**, University of Milano-Bicocca, Italy
Co-authors: Alessandro Colombi, Lucia Paci, Raffaele Argiento

During the last decade, the Bayesian nonparametric community has focused on the definition and investigation of prior distributions in presence of multiple-sample information. A large variety of available models are typically defined by relying on suitable transformations of infinite point processes. Here we define a vector of dependent random probability measures for data organized in groups by normalizing a class of dependent finite point processes. In order to allow the borrowing of information across the diverse groups, we assume that the random probability measures share the same atoms but with different weights. We are able to study all the theoretical properties of the model, i.e., the predictive, posterior and marginal distributions. The random vector of probability measures we propose is then used as a latent structure to define a level-dependent mixture model for clustering with a prior on the number of components. We develop both marginal and conditional algorithms to carry out posterior inference. The performance of the model is tested on several simulated scenarios, and the method is applied to cluster track and field athletes based on their average seasonal performance.

C0408: **Adaptive latent feature sharing for piecewise lineardimensionality reduction**
Presenter: **Yordan Raykov**, University of Nottingham, United Kingdom

Linear Gaussian exploratory tools like PCA and FA are widely used for data analysis and visualization. However, their limitations in high-dimensional problems have led to the development of more robust and flexible models. Discrete-continuous latent feature models offer a solution by inferring the likelihood of shared features among data points. We propose a new approach based on two-parameter discrete distribution models that decouple feature sparsity and dictionary size, capturing common and rare features effectively. This framework enables the development of adaptive variants of factor analysis (aFA) and probabilistic principal component analysis (aPPCA), allowing for flexible structure discovery and dimensionality reduction. We provide efficient inference methods using Gibbs sampling and expectation-maximization, converging much faster to accurate estimates. The effectiveness of aPPCA and aFA is demonstrated in feature learning, data visualization, and data whitening tasks. These models extract meaningful features from MNIST, COLI-20 images, and autoencoder features. Moreover, replacing PCA with aPPCA in functional magnetic resonance imaging (fMRI) analysis improves blind source separation of neural activity, offering more robust and localized results.

CO019 Room BCB 310 DATA DEPTH: A FOCUS ON COMPUTATION AND ANOMALY DETECTION Chair: Pavlo Mozharovskyi

C0258: **Mean estimation, differential privacy, and the sum of squares method**
Presenter: **Sam Hopkins**, Massachusetts Institute of Technology, United States

Recent advances in computational and sample-efficient algorithms are discussed for high-dimensional parameter estimation subject to differential privacy. In particular, several recent works use semidefinite programming relaxations of data-depth problems (via the powerful sum of squares method) together with polynomial-time log-concave sampling algorithms to obtain efficiently-computable estimators for the mean of a high-dimensional distribution with information-theoretically optimal sample complexity. We will discuss these recent results and give an overview of the novel algorithmic techniques underlying them.

C0231: **An efficient algorithm for computing the angular halfspace depth of a whole sample**
Presenter: **Rainer Dyckerhoff**, University of Cologne, Germany
Co-authors: Stanislav Nagy

A great deal of research has recently focused on directional data, i.e., data on the unit sphere. The angular halfspace depth is a tool for nonparametric analysis of directional data. This depth was proposed as early as 1987, but its widespread use in practice has been hampered by significant computational issues. We present an efficient algorithm for exactly computing the angular halfspace depth in arbitrary dimensions. Moreover, this algorithm does not require the data to be in a general position. The algorithm is based on a two-step projection scheme. In the first step, the data is repeatedly projected onto a lower-dimensional sphere. Then, the data is projected from this lower dimensional sphere onto a linear space in which the usual halfspace depth is calculated with respect to a signed measure. Compared to known algorithms, this new algorithm is considerably faster. However, the main advantage of the proposed algorithm is that it is able to compute the depth of all data points in a sample (with respect to that sample) with the same time complexity as the depth of a single point. This is particularly important since calculating the depth of all points in a sample is a common task when using depth-based methods.

C0252: **Exact computation of the scatter halfspace depth**
Presenter: **Stanislav Nagy**, Charles University, Czech Republic

The scatter halfspace depth (sHD) is an extension of the location halfspace (also called Tukey) depth applicable in the nonparametric analysis of scatter. Using sHD, it is possible to define minimax optimal robust scatter estimators for multivariate data. The problem of exact computation of sHD for data of dimension $d > 1$ has, however, not been addressed in the literature. We develop an exact algorithm for the computation of sHD in any dimension d.

C0251: **Affine-invariant integrated rank-weighted depth: Definition, properties and finite sample analysis**
Presenter: **Guillaume Staerman**, Inria, Universita Paris-Saclay, France

Whereas many depth functions have been proposed ad-hoc in the literature since the seminal contribution, not all of them possess the properties desirable to emulate the notion of quantile function for univariate probability distributions. We propose an extension of the integrated rank-weighted (IRW) statistical depth, modified in order to satisfy the property of affine-invariance, fulfilling thus all the four key axioms. The variant we propose, referred to as the Affine-Invariant IRW depth (AI-IRW in short), involves the precision matrix of the d-dimensional random vector X under study,

in order to take into account the directions along which X is the most variable to assign a depth value to any point x in R^d. The accuracy of the sampling version of the AI-IRW depth is investigated from a non-asymptotic perspective. Namely, a concentration result for the statistical counterpart of the AI-IRW depth is proved. Beyond the theoretical analysis carried out, applications to anomaly detection are considered, and numerical results are displayed, providing strong empirical evidence of the relevance of the depth function we propose here.

CO027 Room BCB 309 RECENT ADVANCES IN STATISTICAL LEARNING Chair: Thierry Chekouo

C0193: **Covariance regression with random forests**
Presenter: **Denis Larocque**, HEC Montreal, Canada
Co-authors: Cansu Alakus, Aurelie Labbe

Capturing the conditional covariances or correlations among the elements of a multivariate response vector based on covariates is important to various fields, including neuroscience, epidemiology and biomedicine. We propose a new method called Covariance Regression with Random Forests (CovRegRF) to estimate the covariance matrix of a multivariate response given a set of covariates, using a random forest framework. Random forest trees are built with a splitting rule specially designed to maximize the difference between the sample covariance matrix estimates of the child nodes. We also propose a significance test for the partial effect of a subset of covariates. We evaluate the performance of the proposed method and significance test through a simulation study which shows that the proposed method provides accurate covariance matrix estimates and that the Type-1 error is well controlled. We also demonstrate an application of the proposed method with a thyroid disease data set. CovRegRF is implemented in a freely available R package on CRAN.

C0248: **Clustering with diversity: A promising approach with the determinantal point process**
Presenter: **Serge Vicente**, McGill University, Canada

A random restart of a given algorithm produces many partitions that can be aggregated to yield a consensus clustering. Ensemble methods have been recognized as more robust approaches for data clustering than single clustering algorithms. However, most current initial sets are generated with center points sampled uniformly at random, which can fail both to ensure diversity and obtain good coverage of all data facets. We propose the use of determinantal point processes or DPPs for the random restart of clustering algorithms based on initial sets of center points, such as k-medoids or k-means. DPPs favor diversity of the center points in initial sets, so that sets with similar points have less chance of being generated than sets with very distinct points. Extensive simulations show that DPPs ensure diversity and obtain good coverage of all data facets, two key properties that make DPPs achieve good performance. Simulations with artificial datasets and applications to real datasets show that determinantal consensus clustering outperforms consensus clusterings which are based on a uniform random sampling of center points. The use of DPPs results in final clustering configurations with higher and less dispersed quality scores, when compared to clustering configurations based on uniform sampling of initial points. DPPs are then a promising approach for improving clustering results.

C0266: **Multi-crop land suitability prediction from remote sensing data using semi-supervised learning**
Presenter: **Ayesha Ali**, University of Guelph, Canada
Co-authors: Amanjot Bhullar, Khurram Nadeem

Land suitability prediction involves predicting the area's crop production potential and limitations. We present a data-driven multi-layer perceptron (MLP) that simultaneously predicts the land suitability of several crops in Canada, including barley, peas, spring wheat, canola, oats, and soy. Available crop yields from 2013-2020 are downscaled to the farm level by masking the district-level crop yield data to focus only on areas where crops are cultivated and leveraging soil-climate-landscape variables obtained from Google Earth Engine for crop yield prediction. This new semi-supervised learning approach can accommodate data from different spatial resolutions and enables training with unlabelled data, and allows for the training of a multi-crop model that can capture the interdependences and correlations between various crops, thereby leading to more accurate predictions. The results of our multi-crop model may inform agricultural planning and be incorporated into cost-benefit analyses.

C0247: **A Bayesian variable selection approach incorporating prior feature ordering and population structures**
Presenter: **Thierry Chekouo**, University of Minnesota, United States

A novel Bayesian variable selection framework is proposed for the identification of important genetic variants associated with Coronary artery disease status. Instead of treating each feature independently as in conventional Bayesian variable selection methods, we propose an innovative prior for the inclusion probabilities of genetic variants that account for their ordering structure. We assume that neighboring variants are more likely to be selected together as they tend to be highly correlated and have similar biological functions. Additionally, we propose to group participating subjects based on underlying population structure and fit separate regressions, so that the regression coefficients can better reflect different disease risks in different population groups. Our approach borrows strength across regression models through an innovative prior inspired by the Markov random fields. The proposed framework can improve variable selection and prediction performances, as demonstrated in the simulation studies. We also apply the proposed framework to the CATHGEN data with binary CAD disease status.

CC086 Room BCB 308 MULTIVARIATE STATISTICS Chair: Ray-Bing Chen

C0188: **Variable contribution analysis in multivariate process monitoring using permutation entropy**
Presenter: **Praise Otito Obanya**, North-West University, South Africa
Co-authors: Roelof Coetzer, Carel Olivier, Tanja Verster

Permutation entropy (PE) - a statistical tool for the measurement of the complexity of a given time series - is used to estimate variable contributions to faults in an industrial process. Two sets of simulated industrial chemical process data, namely fault-free and faulty processes, are monitored using Hotelling's T-squared monitoring statistic. For the faults identified, PE is used to determine which variables contributed to those specific faults. Comparisons between the dynamics of the fault-free and faulty processes aid in the identification of the variables with the highest contribution to the specific faults. The well-known Tennessee Eastman Process is used to illustrate the application of PE for variable contributions. The results show that PE is an efficient analysis tool for estimating variable contributions to faults.

C0172: **Optimizing allocation rules in discrete and continuous discriminant analysis**
Presenter: **Dario Ferreira**, University of Beira Interior, Portugal
Co-authors: Sandra Ferreira

An approach is presented to obtain confidence ellipsoids for the vector of probabilities of getting particular results in an experiment with several possible outcomes. We further discuss how to obtain optimal allocation rules, in order to reduce the allocation costs, for both discrete and continuous discriminant analysis. The effectiveness of our approach is demonstrated with two numerical applications. One of them uses real data, while the other one uses simulated data.

C0320: **Variable selection via information gain**
Presenter: **Ting-Li Chen**, Academia Sinica, Taiwan

Variable selection is a critical step in building statistical models, as it helps to identify the most important predictors for explaining the response variable. We will focus on an information gain-based variable selection technique that can effectively handle non-linear relationships between the response and the covariates. We will introduce the information gain criterion, which compares the entropy of the response variable to the

conditional entropy of the response variable given an explanatory variable, to identify the variables that best explain the response. Finally, we will demonstrate the strength of this technique through examples, and compare it with other variable selection methods.

C0365: **Variable discretization-based screening for high dimensional data**
Presenter: **Ryosuke Motegi**, Gunma University, Japan
Co-authors: Yoichi Seki

In statistical modeling with a vast number of predictors, where the number of predictors can be an exponential order of sample size, variable selection plays an essential role in improving prediction performance and interpretability. In such situations, the analyst often takes a two-step process; removing irrelevant predictors with a response variable, then building a model with the remaining predictors using lasso or others. The former step is called variable screening, and various relevance measures have been proposed. An outlier-robust screening method is proposed that can discover variable pairs with nonlinear relationships by looking at the distribution of a response variable conditioned on the discretized predictors. The effectiveness of the proposed method is examined using simulation and real data.

CC110 Room BCB 311 TIME SERIES IN APPLICATIONS Chair: Niklas Ahlgren

C0330: **Multivariate analysis of mortality data using time-varying copula state space models**
Presenter: **Ariane Hanebeck**, Technical University of Munich, Germany
Co-authors: Claudia Czado

The aim is to model and quantify the dependencies between five causes of death, conditional on the weekly number of Covid deaths. Based on the given time series data, we propose to use the model class of copula state space models. The associated latent variable, which we assume to be independent of the number of Covid deaths, can be interpreted as a general driving factor of the causes of death. The dependence between the causes and the latent state however is modeled as varying with the number of Covid deaths. Using this approach, the data in the pre-Covid and the post-Covid time can be modeled within one setup. This leads to a very flexible model allowing for the time dynamics between the causes of death. For the inference, a Bayesian approach is chosen. Due to the high nonlinearity and non-Gaussianity, a Hamiltonian Monte Carlo algorithm is used to sample from the posterior density. After fitting the model, we are able to conduct scenario-based projections for future mortality levels and life expectancy.

C0347: **Feature extraction from satellite data for multivariate time-series forecasting of biotoxin contamination in shellfish**
Presenter: **Sergio Tavares**, NOVA Laboratory for Computer Science and Informatics - NOVA LINCS / NOVA School of Science and Technology, Portugal
Co-authors: Ludwig Krippahl, Marta Lopes

Shellfish production is an important economic activity in Portugal, making shellfish contamination with biotoxins a public health problem and a significant economic risk. Predicting shellfish contamination could improve production management and public health protection. Several years of satellite images obtained from Sentinel-3 mission for marine observation and biotoxin contamination data from shellfish species collected by the Portuguese official control from 16 locations on the western coast of Portugal are used. The goal is to predict when toxin concentration in shellfish will exceed safety limits. The problem is formulated as a time-series forecasting problem, taking as variable past values of contamination and a time series of satellite images for the given locations. Images are available several times a week, while measurements take place once a week. Since images are high-dimensional data, first, a small number of relevant features must be extracted. We do this in an unsupervised manner using autoencoders that are also capable of ignoring non-valid pixels. These frequently occur due to clouds, land or different anomalies. Our results show that including these features improves the prediction of contamination events for 2021 in models trained on data from previous years, showing that with this approach, we can include information from a high-dimension data source like remote sensing without losing the ability of the model to generalize outside the training set.

C0377: **A periodic integer-valued time series with an application to fire activity**
Presenter: **Claudia Susana Santos**, Polytechnic Institute of Coimbra, Portugal
Co-authors: Isabel Pereira

Many INteger-valued AutoRegressive models (INAR) are based on the binomial thinning operator to model non-negative integer-valued time series. In real-life events, data with positive and negative integer values can arise. The signed thinning operator allows for negative values both for the series and its autocorrelation function. Focus is placed upon an INAR(1) process with periodic structure and Skellam-distributed innovations. A brief summary of the definition and properties of the signed periodic INAR model of order 1, denoted by S-PINAR(1), is provided. Forecasting is an important topic in time series. Therefore, point forecasts are computed. A simulation study is conducted to give additional insight into the sample behavior of the forecasts. Approximate point forecasts based on the Gaussian approximation are derived and compared. A performance analysis of prediction intervals for forecasts is presented. We conclude with an application concerning fire activity in Portugal.

C0393: **Nonlinear factor analysis for large sets of macroeconomic time series**
Presenter: **Vivaldo Mendes**, ISCTE-IUL, Portugal
Co-authors: Diana Mendes

Dynamic factor models are frequently used for empirical research in macroeconomics. Several papers in this area have argued that the co-movements of large panels of macroeconomic and financial data can be captured by a relatively few common unobserved (latent) factors. The main purpose of this paper is to analyze and compare the transmission mechanism of monetary policy in the USA, by using a FAVAR (Factor Augmented Vector Auto Regression) model based on two different approaches, in order to include information from large data sets. The first approach consists of a classical linear methodology where the factors are obtained through a principal component analysis (PCA) and Granger causality, while the second one employs a nonlinear factor algorithm based on independent component analysis (ICA), nonlinear PCA, and cross-entropy. In comparison to PCA, the factors extracted by nonlinear methods provide a better performance in the Factor Augmented VAR model, which can be illustrated by Impulse Response Functions and forecasting. We perform the dynamic inference on the model by using a dynamic Bayesian estimation.

Wednesday 23.08.2023 09:00 - 10:30 Parallel Session F – COMPSTAT2023

CI006 Room Virtual room R01 MODERN STATISTICAL ANALYSIS FOR DEPENDENT DATA Chair: Mike So

C0201: **Model averaging for high-dimensional linear regression models with dependent observations**
Presenter: **Henghsiu Tsai**, Academia Sinica, Taiwan

The orthogonal greedy algorithm (OGA) is introduced to screen out the nested set of signal variables under a high-dimensional linear regression framework with dependent observations. To gain the prediction performance, we propose the high-dimensional Mallow model averaging (HDMMA) criteria to determine the weight for averaging these nested high-dimensional linear regression models. We further analyze rates of convergence of prediction error for the averaging model under different sparsity conditions. The contribution has three folds. First, we show that the procedure, named OGA+HDMMA, can achieve optimal convergence rates of prediction error. Second, we use simulation to show that the out-sample prediction of OGA+HAMMA can outperform the MCV method when the covariates are highly correlated or contain time-series effects. Third, the out-sample prediction of OGA+HDMMA performs comparably or even better than many well-known high-dimensional variable selection methods in some scenarios.

C0206: **Dynamic network Poisson autoregression with an application to Covid-19 count data**
Presenter: **Manabu Asai**, Soka University, Japan
Co-authors: Amanda Chu, Mike So

There is a growing interest in accommodating network structure for panel data models. We consider dynamic network Poisson autoregressive (DN-PAR) models for panel count data, allowing time-varying network structure. We develop a Bayesian Markov chain Monte Carlo technique for estimating the DN-PAR model, and we conduct Monte Carlo experiments for examining the property of the posterior quantities to compare dynamic and constant network model. The Monte Carlo results indicate that the bias for the DN-PAR models is negligible, while the constant network model suffers from the bias when the true network is dynamic. We also suggest an approach for extracting the time-varying network from the data. The empirical results for the count data for the confirmed cases of Covid-19 in the United States indicate that the true and extracted dynamic network models outperform the constant network models regarding the deviance information criterion and out-of-sample forecasting.

C0292: **Re-balancing hedge position with statistics of hedge ratios: Concepts and applications**
Presenter: **Cy Sin**, National Tsing Hua University, Taiwan

In a recent article, it has been concluded that "..... there is strong evidence that these models (advanced econometric models) do not improve hedge efficiency significantly, if at all". As a matter of fact, dynamic hedging attempts to strike a balance between hedging effectiveness and transaction costs. Using the Garch asymptotic theories, we derive the asymptotic properties of the hedge ratio. As a result, we construct a natural and simple statistic of re-balancing, namely, the (asymptotic) standard deviation of the hedge ratio. We apply our method to a number of paired variables, such as WTI Crude Oil Futures and Spot Price. Empirical results are compared with those obtained previously.

CO024 Room BCB 308 MULTIDIMENSIONAL VISUALISATION IN ACTION: ADVANCES AND APPLICATIONS Chair: Raeesa Ganey

C0181: **High-dimensional LDA biplot through the GSVD**
Presenter: **Raeesa Ganey**, University of Witwatersrand, South Africa

Linear discriminant Analysis is a multivariate technique concerned with separating distinct sets of observations. However, a common limitation of trace optimisation in discriminant analysis is that the within-cluster scatter matrix must be nonsingular, which restricts the use of data sets when the number of variables is larger than the number of observations. The same goal of discriminant analysis can be achieved by applying the generalised singular value decomposition (GSVD) regardless of the number of variables. We present that by using this approach, we can easily apply discriminant analysis and construct graphical representations to such data. We will look at Canonical Variate Analysis (CVA) biplots that will display observations as points and variables as axes in a reduced dimension, providing a highly informative visual display of the respective class separations.

C0234: **Have house prices factored in the risks of climate change?**
Presenter: **Patricia Menendez**, Monash University, Australia
Co-authors: Maria Jesus Barcena, Maria Cristina Gonzalez, Fernando Tusell

Owning a home is often considered one of the most significant and valuable assets for households, comprising a substantial portion of their net worth. However, the latest report by the Intergovernmental Panel on Climate Change warns of the accelerating global mean sea level rise, which could potentially affect homes in coastal areas and near fluvial regions in the Basque Country. The purpose is to investigate whether house prices have taken into account the risk of flooding associated with climate change. Our dataset comprises more than 300,000 observations, including dwelling transaction prices, rental values, and property characteristics. We use geographically weighted regression with tailored neighbourhoods to analyse the differences in house prices between high-risk and low-risk areas. We use visual representations of data and risk areas to facilitate our analysis and enhance our interpretation of results.

C0242: **Visualizing departures from symmetry: A study on cardiovascular risk among patients with diabetes**
Presenter: **Rosaria Lombardo**, University of Campania, Italy
Co-authors: Eric Beh, Antonio Ceriello, Giuseppe Lucisano, Francesco Prattichizzo, Antonio Nicolucci

Sometimes, the same categorical variable is studied over different time periods or across different cohorts at the same time. In the case of studying the symmetry of a variable, Bowker's chi-squared statistic, presented in 1948, provides a simple numerical means of assessing symmetry for squared contingency tables. We analyse how row and column categories observed on different occasions deviate from the null hypothesis of perfect symmetry. In doing so, a generalization of Bowker's statistic for three-way squared contingency tables is provided. We focus our attention on studying the asymmetry that exists between glycated haemoglobin (HbA1C) variability (in subgroups of subjects with Type 2 diabetes who have a mean HbA1c <53 mmol/mol and > 53 mmol/mol) and cardiovascular complications (myocardial infarction, stroke, coronary revascularization/reperfusion procedures, peripheral revascularization procedures) when additional information is available on the subjects (such as their gender, weight and cholesterol level). This examination of the asymmetry shows the importance of investigating how similar/different, the cardiovascular complications are in cohorts of subjects with varying HbA1C levels that are observed at different occasions. We present a method of analysing and visualizing any departures from symmetry using a variant of correspondence analysis where Bowkers chi-squared statistic is used as its numerical foundation.

C0310: **Explainable outlier detection based on Shapley values**
Presenter: **Marcus Mayrhofer**, TU Wien, Austria
Co-authors: Peter Filzmoser

Shapley values are a popular method in Explainable AI used to quantify the contribution of each feature to the model outcome for each observation. Recently, their practicality has been demonstrated in the context of multivariate outlier detection to explain why an observation deviates from the majority of the data. The Shapley values are used to decompose the squared Mahalanobis distance into outlyingness scores for each variable, which can be interpreted as average marginal contributions to the outlyingness of an observation. The additivity property of the Shapley value ensures that

the sum of those contributions is identical to the squared Mahalanobis distance. Although the computational complexity of the Shapley value is a drawback in general, it can be reformulated and simplified in our case. This reformulation enables quick and efficient computation even in higher dimensions. Multivariate anomaly explanation can be extended to the setting of matrix-variate observations and provides a different perspective on cellwise outlyingness, which aims to detect outlying cells within a data matrix.

CC033 Room BCB 310 ALGORITHMS AND COMPUTATIONAL METHODS Chair: Dominik Liebl

C0296: **Anomaly component analysis: Visualization and interpretability for anomaly detection**
Presenter: **Romain Valla**, Telecom Paris, Institut Polytechnique de Paris, France
Co-authors: Pavlo Mozharovskyi, Florence d Alche-Buc

At the crossway of Machine Learning and Data Analysis, anomaly detection aims at identifying observations that exhibit abnormal behaviour. Be it measurement errors, disease development, severe weather, production quality default(s) (items) or failed equipment, financial frauds or crisis events, their on-time identification and isolation constitute an important task in almost any area of industry and science. While a substantial body of literature is devoted to the detection of anomalies, little attention is paid to their explanation. This is the case mostly due to the intrinsically non-supervised nature of the task and the non-robustness of the exploratory methods like the principal component analysis (PCA). We introduce a new statistical tool dedicated to the exploratory analysis of abnormal observations using data depth as a score. Anomaly component analysis (shortly ACA) is a method that searches a low-dimensional data representation that best visualises and explains anomalies. Based on this, we further propose a procedure for finding clusters of anomalies in Euclidean space. In a comparative study using simulated and real data, ACA proves advantageous for anomaly analysis with respect to methods present in the literature.

C0298: **A simple and direct procedure for data generation in PLS-SEM framework**
Presenter: **Sophie Dominique**, XLSTAT, France
Co-authors: Veronique Cariou, Mohamed Hanafi, Jean-Marc Ferrandi, Fabien Llobell

Simulation studies usually require an initial step of generating datasets by varying the values of several pre-defined parameters. As with all statistical techniques, it is necessary to carry out simulation studies within the composite-based SEM framework to check the validity conditions, compare the performance and reliability of related methods, or identify the most efficient one under certain assumptions. Among the different composite-based SEM methods already proposed, we focus here on the PLS-SEM method, which remains undeniably the most widely used. While data generation procedures have been extensively studied in the context of covariance-based approaches, little attention has been paid to adapting these procedures to the context of composite-based approaches and, more specifically, to PLS-SEM. To fill the gap, we propose a direct data generation procedure that follows the logic of the PLS-SEM algorithm and involves the linear regressions derived from the structural model to estimate path coefficients. This procedure turns out to be simpler than the one recently proposed by Schlittgen. The main reason is that Schlittgen's procedure requires the calculation of the covariance matrix implied by the structural model, whereas such a calculation is unnecessary for our new data generation scheme. Finally, a comparison is made between the two different strategies evaluating the accuracy of the parameter values obtained with respect to those defined a priori.

C0302: **dataSDA and ggESDA: Two R packages for exploratory symbolic data analysis**
Presenter: **Han-Ming Wu**, National Chengchi University, Taiwan

Exploratory Data Analysis (EDA) serves as a preliminary yet essential tool for summarizing the main characteristics of a dataset before appropriate statistical modeling can be applied. Quite often, EDA employs traditional graphical techniques such as boxplots, histograms, and scatterplots, and is equipped with various dimension reduction methods and computer-aided interactive functionalities. Over the years, data collected has become increasingly large and complex. Data descriptions have moved beyond single-value representations, encompassing intervals, histograms, and distributions. These are examples of the so-called symbolic data. In response to this development, we have created two R packages: dataSDA and ggESDA. The dataSDA package is designed to collect a diverse range of symbolic data and offers a comprehensive set of functions that facilitate the conversion of traditional data into the symbolic data format. These datasets can serve as benchmarks for evaluating symbolic data analysis methods. In addition, the package implements various R functions for computing symbolic descriptive statistics. The ggESDA package extends ggplot2 to offer a variety of plots specifically designed for exploratory symbolic data analysis. We will discuss how ggESDA is implemented. We will demonstrate its utility through the analysis of two real symbolic datasets found in dataSDA.

CC057 Room BCB 311 SEMI- AND NONPARAMETRIC METHODS Chair: Ivan Kojadinovic

C0321: **Prediction of order statistics based on ordered generalized ranked set sampling**
Presenter: **Masato Kitani**, Tokyo University of Science, Japan
Co-authors: Katsuyuki Yuasa, Hidetoshi Murakami

In statistical inference, a prediction interval is important for estimating an interval in which a future observation will fall. The distribution-free prediction intervals for order statistics of future observations have been introduced previously. In practical analysis, the prediction of future extreme observations plays an important role, such as the largest river flow in the next few years. However, large sample sizes are needed to construct the prediction intervals with sufficient coverage probability for future extreme order statistics. Therefore, we propose a prediction interval using the generalized ranked set sampling, which can observe only selected order statistics by using the rank of the data before observation. We show that the proposed prediction interval has sufficient coverage probabilities with small sample sizes. Furthermore, we show that the proposed prediction interval has desirable properties compared to other existing methods by numerical simulations.

C0334: **A measure for the degree of distribution changes in locally stationary processes**
Presenter: **Guy-Niklas Brunotte**, Otto-Friedrich-Universitat Bamberg, Germany

Non-stationary time series $\{X_{t,T}\}$ with $t = 1,\dots,T$ and $T \in \mathbb{N}$ are considered in many applications. Thereby, it is often fundamental to know how much the distribution of $X_{t,T}$ depends on the concrete point in time t because such an analysis contributes to answering how good estimators based on historical data forecast critical values concerning the future. For example, it is of interest whether the distributions of daily returns of a stock are more time-dependent after a crisis than before this event or even stay the same for all investigated points in time. This motivates the introduction of a characteristic function-based, well-interpretable measure which quantifies for a wide class of non-stationary processes $\{X_{t,T}\}$ how much the distribution of $X_{t,T}$ depends on t. Moreover, it is shown that this measure provides an asymptotic level alpha test which examines whether the distribution of $X_{t,T}$ changes over time. In addition, the present measure will be applied to several daily returns of stocks to investigate how much their distributions depend on time.

C0372: **Bandwidth selection method for estimating difference between two densities with kernel density estimation**
Presenter: **Sixiao Zhu**, Paris 1 University, France
Co-authors: Alain Celisse

Measuring the difference between two probability distributions is the key issue of many statistical applications such as change point detection and clustering. Calculating the L2 distance between two kernel estimations of densities serves a straight forward measure of this kind. This brings forward the classical issue of choosing the bandwidth of kernel used in this procedure. In this work, we address this issue by borrowing idea from the recent work of penalized comparison to overfitting (PCO) method aiming originally the density estimation problem, and propose a new

bandwidth selection method for our two-sample setting. We justify the performance of the proposed method by establishing theoretical results such as oracle inequality, together with numerical simulation results.

C0207: **Open-end monitoring for multivariate observations sensitive to all types of changes in the distribution function**
Presenter: **Ivan Kojadinovic**, CNRS UMR 5142 LMA University of Pau, France
Co-authors: Mark Holmes, Alex Verhoijsen

Nonparametric open-end sequential testing procedures are proposed based on the empirical distribution function that can detect all types of changes in the contemporary distribution function of multivariate observations. Their asymptotic properties are theoretically investigated under stationarity and under alternatives to stationarity. Monte Carlo experiments reveal their good finite-sample behavior in the case of continuous low-dimensional observations. A short data example concludes the presentation.

CC046 Room BCB 309 MACHINE LEARNING Chair: Peter Winker

C0204: **Multiclass machine learning classification of functional brain images for Parkinson's disease stage prediction**
Presenter: **Guan-Hua Huang**, National Yang Ming Chiao Tung University, Taiwan

A data set is analyzed that contains functional brain images from 6 healthy controls and 196 individuals with Parkinson's disease (PD), who were divided into five stages according to illness severity. The goal was to predict patients' PD illness stages by using their functional brain images. We employed the following prediction approaches: multivariate statistical methods (linear discriminant analysis, support vector machine, decision tree, and multilayer perceptron [MLP]), ensemble learning models (random forest [RF] and adaptive boosting), and deep convolutional neural network (CNN). For statistical and ensemble models, principal component analysis was performed to extract features, and synthetic minority over-sampling technique (SMOTE) was used to deal with imbalanced data problems. For CNN modeling, we applied an image augmentation technique to increase and balance data sizes over different disease stages, and we adopted a transfer learning idea to bring pre-trained VGG16 weights and architecture into the model fitting. It was found that MLP and RF were the analytic approaches with the highest prediction accuracy rate for statistical and ensemble models, respectively. Overall, the deep CNN model with pre trained VGG16 weights and architecture outperformed other approaches.

C0217: **Two-sample testing in reinforcement learning**
Presenter: **Martin Waltz**, TU Dresden, Germany
Co-authors: Ostap Okhrin

Value-based reinforcement-learning algorithms have shown strong performances in games, robotics, and other real-world applications. The most popular sample-based method is Q-Learning. It subsequently performs updates by adjusting the current Q-estimate towards the observed reward and the maximum of the Q-estimates of the next state. The procedure introduces maximization bias with approaches like Double Q-Learning. We frame the bias problem statistically and consider it an instance of estimating the maximum expected value (MEV) of a set of random variables. We propose the T-Estimator (TE) based on two-sample testing for the mean, that flexibly interpolates between over- and underestimation by adjusting the significance level of the underlying hypothesis tests. A generalization, termed K-Estimator (KE), obeys the same bias and variance bounds as the TE while relying on a nearly arbitrary kernel function. We introduce modifications of Q-Learning and the Bootstrapped Deep Q-Network (BDQN) using the TE and the KE. Furthermore, we propose an adaptive variant of the TE-based BDQN that dynamically adjusts the significance level to minimize the absolute estimation bias. All proposed estimators and algorithms are thoroughly tested and validated on diverse tasks and environments, illustrating the bias control and performance potential of the TE and KE.

C0361: **Generalized additive models for multiclass detection of voice disorders by using acoustic features**
Presenter: **Carlos Javier Perez Sanchez**, University of Extremadura, Spain
Co-authors: Lizbeth Naranjo, Victor Miranda, Josefa Hernandez

Computer-aided diagnosis systems for detecting voice-related diseases from speech recordings require developing and using reliable statistical models. In a binary classification context, several approaches have addressed the problem of discriminating between healthy and pathological subjects based on acoustic features extracted from voice recordings. In these cases, the binary classification was used to distinguish between healthy and single pathology subjects or between healthy and a group of subjects with different pathologies. However, the multiclass problem has not been sufficiently addressed, as there is difficulty in discriminating between two voice pathologies. A generalized additive model with a variable selection procedure has been implemented and applied to classify subjects with vocal fold nodules, Reinkes edema, and without any pathology. An in-house built database of voice recording based on the sustained phonation of the vowel was used. A total of 30 acoustic features were extracted using various speech processing algorithms, covering a wide variety of feature types, including nonlinear ones. The approach achieved an accuracy of 93%, compared to 71% accuracy obtained with logistic regression. This highlights the potential value of the nonlinear nature of the generalized additive model in addressing multiclass problems in this context.

C0369: **On the ability of random forests to model interactions**
Presenter: **Constanze Lehner**, University of Passau, Germany

It is often argued that random forests can implicitly account for interactions due to the tree-like structure of its base learner. The binary recursive partitioning approach of classification and regression trees divides the observations into more homogeneous subgroups by making sequential univariate splits on covariates. Recent contributions suggest that such univariate partitioning schemes may not be appropriate to adequately capture the effect of interrelated covariates on the response variable. To explore these arguments about whether interactions are modeled in random forests, we use simulated datasets with different interaction patterns and evaluate the predictive performance and variable importance measures of random forests. Our analysis also provides a review of interactions in the random forest literature, from their origins in the decision tree literature to current applications. For the review, we discuss interaction concepts in different research areas in advance.

Wednesday 23.08.2023 11:00 - 12:30 Parallel Session G – COMPSTAT2023

CI004 Room BCB 307 CHANGE-POINT ANALYSIS Chair: Ivan Kojadinovic

C0156: **Data segmentation: Moving-sum-procedures and bootstrap confidence intervals**
Presenter: **Claudia Kirch**, Otto-von-Guericke University Magdeburg, Germany
Co-authors: Haeran Cho

Using the example of changes in the mean, we introduce change point estimators for multiple changes based on moving sum statistics which obtain optimal localisation rates for the change points. Despite the optimality property, these estimators are not consistent in the usual sense but instead exhibit uncertainty that is non-vanishing, even asymptotically in non-rescaled time. We propose bootstrap estimators to tame this uncertainty by means of confidence intervals and show that the bootstrap automatically adapts to the different asymptotic regimes associated with local and fixed changes, respectively. We shortly discuss extensions to multiple bandwidths moving sum procedures which work well in the presence of multiscale change points with both large jumps over short intervals and small changes over long stationary intervals.

C0199: **Sparse change detection in high-dimensional linear regression**
Presenter: **Tengyao Wang**, London School of Economics, United Kingdom

A new methodology 'charcoal' is introduced for estimating the location of sparse changes in high-dimensional linear regression coefficients, without assuming that those coefficients are individually sparse. The procedure works by constructing different sketches (projections) of the design matrix at each time point, where consecutive projection matrices differ in sign in exactly one column. The sequence of sketched design matrices is then compared against a single sketched response vector to form a sequence of test statistics whose behaviour shows a surprising link to the well-known CUSUM statistics of univariate changepoint analysis. Strong theoretical guarantees are derived for the estimation accuracy of the procedure, which is computationally attractive, and simulations confirm that our methods perform well in a broad class of settings.

C0240: **Change point analysis of functional time series**
Presenter: **Gregory Rice**, University of Waterloo, Canada
Co-authors: Alexander Aue, Lajos Horvath, Yuqian Zhao, Jeremy Vander Does, Ozan Sonmez

Several recent advances in change point analysis for functional time series are surveyed. Beginning with simple methods to detect and estimate changes in the mean function, we go on to consider change point methods for second-order and higher-order properties of a functional time series. These methods are illustrated with applications time series of curves derived from the energy market and environmental data. We conclude with a discussion of open avenues of research in the area.

CO100 Room Virtual room R01 CLUSTERING AND REGRESSION ANALYSIS OF COMPLEX REAL-LIFE DATA Chair: Gabriele Soffritti

C0218: **Mixture models for heavy-tailed tensor-variate data**
Presenter: **Salvatore Daniele Tomarchio**, University of Catania, Italy
Co-authors: Antonio Punzo, Luca Bagnato

Real data is assuming increasingly complicated structures, requiring more flexible statistical approaches. Tensor-variate (or multi-way) structures are a typical example of such kind of data. Unfortunately, atypical observations often occur in real data, making the traditional normality assumption inadequate. To cope with this problem, we introduce two tensor-variate distributions, both heavy-tailed generalizations of the tensor-variate normal distribution. Then, using finite mixture models, we use these distributions for model-based clustering. We apply the eigen-decomposition of the components' scale matrices to reach parsimony, resulting in two families of parsimonious tensor-variate mixtures. We illustrate variants of the well-known EM algorithm for parameter estimation. Since the order of the tensors affects the number of parsimonious models, we implement strategies intending to shorten the initialization and fitting processes. Simulated analyses are used to study these processes. Last but not least, we applied our parsimonious models to real datasets.

C0261: **Non-parametric multi-partitions clustering**
Presenter: **Vincent Vandewalle**, Universite Cote d'Azur, France
Co-authors: Marie du Roy de Chaumaray

In the framework of model-based clustering, a model, called multi-partitions clustering, allowing several latent class variables has been proposed. This model assumes that the distribution of the observed data can be factorized into several independent blocks of variables, each block following its own mixture model. We assume that each block follows a non-parametric latent class model, i.e., independence of the variables in each component of the mixture with no parametric assumption on their class conditional distribution. The purpose is to deduce, from the observation of a sample, the number of blocks, the partition of the variables into the blocks and the number of components in each block, which characterise the proposed model. By following recent literature on model and variable selection in non-parametric mixture models, we propose to discretize the data into bins. This permits to apply the classical multi-partition clustering procedure for parametric multinomials, which are based on a penalized likelihood method (e.g., BIC). The consistency of the procedure is obtained, and an efficient optimization is proposed. The performances of the model are investigated on simulated data.

C0272: **Model based labelling of hyperspectral food images**
Presenter: **Ganesh Babu**, University College Dublin, Ireland

Food engineers have been using hyperspectral images to classify the type and quality of a food sample under study, typically using machine learning (ML) classifiers. In order to train these classifiers, threshold-based approaches are used to label the pixels and classifiers are then trained based on those labels. However, the threshold-based approaches are ad-hoc, subjective and cannot be generalized across hyperspectral images. To address this issue, a consensus-constrained parsimonious Gaussian mixture model (ccPGMM) is proposed to cluster the pixels of the hyperspectral images. The resulting cluster labels can then be employed for training the classifiers in the classification phase. The ccPGMM utilizes visual and spatial information from appropriately selected subsets of pixels. This information is used as a constraint when clustering the rest of the pixels in the image. In this way, the model ensures that pixels in the same areas of the image are clustered together, while pixels located in different regions of the image are not allocated to the same cluster. The hyperspectral images are high-dimensional - a parsimonious latent variable model and a consensus clustering approach are employed to handle this. With the consensus approach, the data are divided into multiple subsets of variables and the constrained clustering is applied to each subset. The clustering results are then combined across all the subsets to provide a consensus clustering solution for each pixel.

C0259: **Mixtures of linear regression models: An application to housing tension in Emilia-Romagna, Italy**
Presenter: **Gabriele Soffritti**, University of Bologna, Italy
Co-authors: Gabriele Perrone

Mixtures of multivariate linear regression models constitute an approach which simultaneously allows performing linear regression analysis and model-based cluster analysis. They are generally employed when sample observations come from a population composed of unknown sub-populations. Extensions of such models have been recently introduced so as to manage the possible presence of mild outliers and let the researcher be free to use a different vector of covariates for each response. As a consequence, complex real-life data from many areas of activity can be

adequately analysed using this type of models. An example is represented by housing deprivation in Italy. As far as the case of the Emilia-Romagna region is concerned, an observatory of the housing system regularly monitors housing conditions and supports the development of public housing policies at a municipality level. Within this framework, a dataset provided by the Emilia-Romagna region has been analysed in order to identify the socio-demographic, income and housing market factors that affect housing tension in the municipalities of the region. The effects of such factors have been studied. Heterogeneity in municipalities has been detected. In order to ensure additional flexibility, the analysis has been complemented by the use of models whose mixing proportions are expressed as functions of some concomitant covariates.

CO025 Room BCB 310 ADVANCES IN STATISTICS FOR FINANCE Chair: Massimiliano Caporin

C0286: **Sparse graphical modelling for minimum variance portfolios**
Presenter: **Giovanni Bonaccolto**, University of Enna Kore, Italy
Co-authors: Riccardo Riccobello, Philipp Johannes Kremer, Sandra Paterlini, Malgorzata Bogdan
Graphical models have demonstrated exceptional performance in uncovering the conditional dependence structure among a given set of variables. Two novel graphical modeling techniques are introduced: Gslope and Tslope, which use the Sorted L1-Penalized Estimator (Slope) to directly estimate the inverse of the covariance matrix. We develop ad hoc algorithms to efficiently solve the underlying optimization problems: the Alternating Direction Method of Multipliers for Gslope and the Expectation-Maximization (EM) algorithm for Tslope. The methods are suitable for both Gaussian and non-Gaussian distributed data and take into account the empirically observed distributional characteristics of asset returns. Through extensive simulation analysis and real-world applications, we demonstrate the superiority of our new methods over state-of-the-art covariance matrix estimation techniques, particularly regarding clustering and stability characteristics.

C0223: **Nonlinear scalar BEKK**
Presenter: **Bilel Sanhaji**, University Paris 8, France
A nonlinear conditional covariance model driven by five scalars is proposed. We propose Lagrange Multiplier tests for nonlinearity in conditional covariances of multivariate GARCH models. We also show asymptotic properties of the tests through Monte Carlo simulations, and provide empirical illustrations.

C0225: **Forecast calibration, backtests, and score decompositions for Value-at-Risk**
Presenter: **Marius Puke**, University of Hohenheim, Germany
Co-authors: Timo Dimitriadis
The evaluation of Value-at-Risk (VaR) forecasts is fundamental to the stability of our financial system through the regulatory frameworks Basel III for banks and Solvency II for insurance, but also their internal risk forecasts. While the statistical literature advises the use of scoring (loss) functions for forecast evaluation, the economic literature, as well as the practical implementation in the regulatory systems, is still mainly concerned with so-called VaR backtests. The fundamental connection between these two principles is still not well understood, which is especially problematic as these approaches regularly deliver contrasting results in practice. We formally connect these concepts by drawing on the recent literature on forecast calibration. For this, we make use of recently developed decompositions of scoring functions into interpretable components assessing (mis)calibration, discrimination, and uncertainty and show that the backtests only assess calibration while entirely ignoring the forecasts discriminating ability. Intuitively, this corresponds to ignoring the forecast's ability to discriminate between periods of lower and higher risk. As a consequence, we propose the practical use of score decompositions that, in addition to backtests, reveal information on the overall predictive ability and the hitherto unexploited information on discrimination and hence provoke more informative insights in applications.

C0174: **Penalized CAW, forecast error variance decompositions and systemic risk measurement**
Presenter: **Massimiliano Caporin**, University of Padova, Italy
Co-authors: Giuseppe Storti
Parameter estimation of the Conditional Autoregressive Wishart model under penalization is discussed. We introduce two novel Forecast Error Variance Decompositions where returns shocks impact on the entire set of realized variances and covariances, the first following a more traditional approach and the second based on simulations. From both decompositions, we derive a spillover index to monitor the systemic risk. An empirical analysis of US large cap equities exemplifies our proposals.

CO008 Room BCB 309 CAUSAL INFERENCE AND FUNCTIONAL DATA ANALYSIS Chair: Nicolas Hernandez

C0191: **Causal inference with functional data**
Presenter: **Dominik Liebl**, University Bonn, Germany
Co-authors: Tim Mensinger
Building upon the potential outcome framework for functional data, a method is proposed for inferring treatment effect functions in function-on-scalar regression models using simultaneous confidence bands based on adaptive critical value functions. The proposed method allows for inference under fairness constraints, enabling both local and global interpretability of the results. We present convergence and asymptotic normality results for the functional version of the doubly-robust augmented inverse propensity score estimator.

C0224: **Causal inference with a functional outcome**
Presenter: **Kreske Ecker**, Umea University, Sweden
Co-authors: Xavier de Luna, Lina Schelin
Methods are presented to study the causal effect of a binary treatment on a functional outcome with observational data. We define a functional causal parameter, the Functional Average Treatment Effect (FATE), and propose a semi-parametric outcome regression estimator. Quantifying the uncertainty in this estimation presents a challenge since existing inferential techniques developed for univariate outcomes cannot satisfactorily address the multiple comparison problems induced by the functional nature of the causal parameter. We show how to obtain valid inferences on the FATE using simultaneous confidence bands, which cover the FATE with a given probability over the entire domain. Simulation experiments illustrate the empirical coverage of the simultaneous confidence bands in finite samples. Finally, we use the methods to infer the effect of early adult location on subsequent income development for one Swedish birth cohort.

C0226: **High-dimensional nonparametric functional graphical models via the functional additive partial correlation operator**
Presenter: **Eftychia Solea**, Queen Mary University of London, United Kingdom
A novel approach is developed for estimating a nonparametric graphical model for functional data. The approach is built on a new linear operator, the functional additive partial correlation operator, which extends the partial correlation matrix to both the nonparametric and functional settings. We establish both estimation consistency and graph selection consistency of the proposed estimator, while allowing the number of nodes to grow with the increasing sample size. Through simulation studies, we demonstrate that our method performs better than existing methods in cases where the Gaussian or Gaussian copula assumption does not hold. We also demonstrate the performance of the proposed method by a study of an electroencephalography data set to construct a brain network.

C0263: **Elastic linear regression for curves in R^d**
Presenter: **Sonja Greven**, Humboldt University of Berlin, Germany
Co-authors: Lisa Steyer, Almond Stoecker

Regression models are proposed for curve-valued responses in two or more dimensions, where only the image but not the parametrisation of the curves is of interest. Examples of such data are handwritten letters, movement paths or outlines of objects. In the square-root-velocity framework, a parametrisation invariant distance for curves is obtained as the quotient space metric with respect to the action of re-parametrisation, which is by isometries. With this special case in mind, we discuss the generalisation of 'linear' regression to quotient spaces more generally, before illustrating the usefulness of our approach for curves modulo re-parametrisation. We address the issue of irregularly sampled curves by using splines for modelling smooth predicted curves. We test this model in simulations and apply it to human hippocampi data, obtained from MRI scans. We model how the shape of the hippocampus is related to age and Alzheimer's disease.

CC082 Room BCB 311 HIGH-DIMENSIONAL STATISTICS Chair: Maria Brigida Ferraro

C0341: **Two-sample test based on the variance of a positive definite kernel**

Presenter: **Natsumi Makigusa**, Chuo University, Japan

The aim is to test whether the distributions P and Q followed by two samples are the same. Especially a two-sample test based on Maximum Mean Discrepancy (MMD) is known as an approach for high-dimensional low-sample size data. The MMD embeds a probability distribution into the reproducing kernel Hilbert space by the mean of the positive definite kernel and measures the difference between the two distributions P and Q based on the distance between each embedding. We introduce a novel discrepancy called the Maximum Variance Discrepancy (MVD) for the purpose of measuring the difference between two distributions in Hilbert spaces that cannot be found via the MMD. The MVD measures the difference between the distributions P and Q by embedding the variances of the definite positive kernel under the P and Q into the tensor product space of the reproducing kernel Hilbert space and measuring these differences. We propose a two-sample test based on this MVD and obtain the asymptotic distributions of this test statistic. The asymptotic null distribution of this test statistic is the infinite sum of the weighted chi-square distribution. We propose an approximation of the null distribution to obtain the critical value.

C0208: **A distribution-free change-point monitoring scheme in high-dimensional settings**

Presenter: **Niladri Chakraborty**, University of the Free State, South Africa

Co-authors: Chun Fai Lui, Maged Ahmed

Existing monitoring tools for multivariate data are often asymptotically distribution-free, computationally intensive, or require a large stretch of stable data. Many of these methods are not applicable to high-dimension, low-sample size scenarios. With rapid technological advancement, high-dimensional data has become omnipresent in industrial applications. We propose a distribution-free change-point monitoring method applicable to high-dimensional data. Through an extensive simulation study, performance comparison has been done for different parameter values, under different multivariate distributions with complex dependence structures. The proposed method is robust and efficient in detecting change points under a wide range of shifts in the process distribution. A real-life application is illustrated with the help of a high-dimensional image surveillance dataset.

C0370: **Skew-normal classification in high-dimensional data**

Presenter: **Haesong Choi**, Florida state university, United States

Co-authors: Qing Mai

A considerable number of studies have been devoted to high-dimensional classification models under the assumption of normality. However, the normality assumption may be too restrictive in data analysis. Motivated by the data sets that exhibit asymmetry, including environmental, financial, and biomedical ones, we propose a high-dimensional discriminant analysis model called the SKNC model (short for SKew-Normal Classification). By incorporating the skew-normal distribution, the SKNC model is closely related to the LDA model but improves its flexibility on skewed data in classification. We further develop a novel classifier to estimate the SKNC model. To tackle the statistical challenge of the heavy-tailed distribution, we propose a robust estimation of parameters. We develop an efficient algorithm to adopt the penalized estimation. Theoretical results rigorously show that the SKNC model achieves variable selection and penalized estimation, especially in high-dimensional settings. We empirically demonstrate the superior performance of the SKNC model over existing methods in simulated and real datasets.

C0333: **Inference for low-rank models without rank estimation**

Presenter: **Hyukjun Kwon**, Rutgers University, United States

Co-authors: Yuan Liao, Jungjun Choi

A new debiasing procedure is introduced for linear low-rank models, where the parameter of interest is a high-dimensional matrix coefficient. Our procedure achieves asymptotic normality without requiring knowledge of the true rank of the parameter matrix. The key feature of our approach is the use of diversified weights. We project an intermediate estimator onto low-rank linear spaces that are estimated using these weights. Notably, this projection is robust to the rank misspecification. However, the estimated projection matrices are inconsistent with the true projections, creating new challenges in characterizing the asymptotic distribution of our debiased estimator. Nonetheless, our proposed debiasing procedure successfully addresses these issues. Lastly, our procedure does not require sample splitting.

Wednesday 23.08.2023 14:15 - 15:45 Parallel Session H – COMPSTAT2023

CV072 Room Virtual room R01 SPATIAL STATISTICS Chair: Ivan Kojadinovic

C0329: **Asymptotic analysis of ML-covariance parameter estimators based on covariance approximations**
Presenter: **Michael Hediger**, University of Zurich, Switzerland
Co-authors: Reinhard Furrer

Given a zero-mean Gaussian random field with a covariance function that belongs to a parametric family of covariance functions, we introduce a new notion of likelihood approximations, termed truncated-likelihood functions. Truncated-likelihood functions are based on direct functional approximations of the presumed family of covariance functions. For compactly supported covariance functions, within an increasing-domain asymptotic framework, we provide sufficient conditions under which consistency and asymptotic normality of estimators based on truncated-likelihood functions are preserved. We apply our result to the family of generalized Wendland covariance functions and discuss several examples of Wendland approximations. For families of covariance functions that are not compactly supported, we combine our results with the covariance tapering approach and show that ML estimators, based on truncated-tapered likelihood functions, asymptotically minimize the Kullback-Leibler divergence, when the taper range is fixed.

C0346: **Fast Bayesian inference of block nearest neighbor Gaussian models for large data**
Presenter: **Zaida Quiroz**, Pontificia Universidad Catolica del Peru, Peru
Co-authors: Marcos Prates, Dipak Dey, Haavard Rue

The purpose is to present the development of a spatial block-nearest neighbor Gaussian process (blockNNGP) for location-referenced large spatial data. The key idea behind this approach is to divide the spatial domain into several blocks, which are dependent on some constraints. The cross-blocks capture the large-scale spatial dependence, while each block captures the small-scale spatial dependence. The resulting blockNNGP enjoys Markov properties reflected on its sparse precision matrix. It is embedded as a prior within the class of latent Gaussian models. Thus, fast Bayesian inference is obtained using the integrated nested Laplace approximation. The performance of the blockNNGP is illustrated on simulated examples, a comparison of our approach with other methods for analyzing large spatial data and applications with Gaussian and non-Gaussian real data.

C0171: **A generalized functional linear model with spatial dependence**
Presenter: **Sooran Kim**, Iowa State University, United States
Co-authors: Mark Kaiser, Xiongtao Dai

A regression model is developed for spatially dependent binary response variables when the covariates take the form of functional processes over time at each location for which the response is observed. We model the functional covariates in terms of a Fourier basis truncated to a finite number of terms. Responses are taken to be a Markov random field with conditional binary distributions and isotropic spatial dependence. Estimation is approached through the use of a composite likelihood constructed from full conditional response distributions, sometimes also called Besags original pseudolikelihood in the spatial literature. Asymptotic properties are given for maximum composite likelihood estimators using a repeating lattice context, and the use of the model is illustrated with data relating new COVID vaccination rates in June for counties to the number of weekly infections reported over the previous several months in those same counties.

CO104 Room BCB 308 STATISTICS FOR DATA SCIENCE Chair: Luis Alberto Firinguetti Limone

C0166: **Estimation of expectations in two-level nested simulation experiments**
Presenter: **David Fernando Munoz**, Instituto Tecnologico Autonomo de Mexico, Mexico

Input parameters of a simulation experiment are usually estimated from real-data observations, and parameter uncertainty can be significant when little data is available. In this case, Bayesian statistics can be used to incorporate this uncertainty in the output analysis of simulation experiments via the use of a posterior distribution. A methodology currently proposed for the analysis of simulation experiments under parameter uncertainty is a two-level nested simulation method. In the outer level, we simulate n observations for the parameters from a posterior distribution, while in the inner level, we simulate m observations for the response variable with the parameter fixed at the value generated in the outer level. In this paper, we focus on the output analysis of two-level simulation experiments, for the case where the observations of the inner level are independent, showing how the variance of a simulated observation can be decomposed into parametric and stochastic variance components. We derive a Central Limit Theorem (CLT) for both the estimator of the point forecast and the estimators of the variance components. Our CLTs allow us to compute asymptotic confidence intervals for each estimator. Our theoretical results are validated through experiments with a forecasting model for sporadic demand.

C0178: **Bayesian spatio-temporal modeling of the Brazilian wildfires: The influence of human and meteorological variables**
Presenter: **Paulo Canas Rodrigues**, Federal University of Bahia, Brazil

Wildfires are among the most common natural disasters in many world regions and actively impact life quality. These events have become frequent due to climate change, other local policies, and human behavior. The historical data with the geographical locations of all the fire spots detected by the reference satellites covering the Brazilian territory between January 2011 and December 2022 are considered, comprising more than 2.2 million fire spots. This data were modeled with a spatio-temporal generalized linear model for areal unit data, whose inferences about its parameters are made in a Bayesian approach and use meteorological variables (precipitation, air temperature, humidity, and wind speed) and a human variable (land-use transition and occupation) as covariates. The change in land use from the forest and green areas to farming significantly impacts the number of fire spots for all six Brazilian biomes. (Joint work with Jonatha Pimentel and Rodrigo Bulhoes)

C0197: **Building and classifying brain images**
Presenter: **Martha Bohorquez**, Universidad Nacional de Colombia, Colombia

Some novel methodologies are presented that combine functional geostatistics and deep learning techniques for building, analyzing and classifying image time series. We focus on pattern recognition in electroencephalography (EEG) data.

C0200: **Shrinkage estimators for beta regression models**
Presenter: **Luis Alberto Firinguetti Limone**, Universidad del Bio-Bio, Chile
Co-authors: Luis Gomez

The beta regression model is a useful framework for studying response variables which are rates or proportions, that is to say, response variables which are continuous and restricted to the interval $(0, 1)$. As with any other regression model, parameter estimates may be affected by collinearity, or even perfect collinearity, among the explanatory variables. To handle these situations, shrinkage estimators are proposed. In particular, we develop Ridge Regression and LASSO estimators from a penalized likelihood perspective with a logit link function. The properties of the resulting estimators are evaluated through simulation experiments.

CO103 Room BCB 310 DYNAMIC NETWORKS Chair: Philip Yu

C0159: **Triangular concordance learning of networks**
Presenter: **Jiaqi Gu**, Stanford University, United States

Co-authors: Guosheng Yin

Networks are widely used to describe relational data among objects in a complex system. As network data often exhibit clustering structures, research interest often focuses on discovering clusters of nodes. We develop a novel concordance-based method for node clustering in networks, where a linear model is imposed on the latent position of each node with respect to a node-specific center and its covariates via linear transformation. By maximizing a triangular concordance function with a concave pairwise penalty, the latent positions are estimated so that each node would be more likely to be close to its neighbors in contrast to non-neighbors and nodes are clustered by their node-specific centers. We develop an alternating direction method of multipliers algorithm for parameter estimation and an intimacy score between unlinked nodes for link prediction. Our method takes into account common characteristics of network data (i.e., assortativity, link pattern similarity, node heterogeneity and link transitivity), while it does not require the number of clusters to be known. The clustering effectiveness and link prediction accuracy of our method are demonstrated in simulated and real networks.

C0189: **A two-way heterogeneity model for dynamic networks**

Presenter: **Binyan Jiang**, The Hong Kong Polytechnic University, Hong Kong

Analysis of networks that evolve dynamically requires the joint modelling of individual snapshots and time dynamics. A new flexible two-way heterogeneity model towards this goal is proposed. The new model equips each node of the network with two heterogeneity parameters, one to characterize the propensity to form ties with other nodes statically and the other to differentiate the tendency to retain existing ties over time. With n observed networks each having p nodes, we develop a new asymptotic theory for the maximum likelihood estimation of $2p$ parameters when $np \to \infty$ in which $n \geq 2$ can be finite. We overcome the global non-convexity of the negative log-likelihood function by virtue of its local convexity, and propose a novel method of moment estimator as the initial value for a simple algorithm that leads to the global maximum likelihood estimator (MLE). To establish the upper bounds for the estimation error of the MLE, we derive a new uniform deviation bound, which is of independent interest. The theory of the model and its usefulness are further supported by extensive simulation and a data analysis examining the social interactions of ants.

C0284: **High-dimensional low-rank linear time series modeling**

Presenter: **Guodong Li**, University of Hong Kong, Hong Kong

Low-rank structures are imposed to the column and row spaces of coefficient matrices in a multivariate infinite-order vector autoregression (VAR), which, with the help of tensor techniques, leads to a newly proposed concept of supervised factor models, where two-factor modelings are conducted to responses and predictors simultaneously. Interestingly, the stationarity condition implies an intrinsic weak group sparsity mechanism of infinite-order VAR, and hence a rank-constrained group Lasso estimation is considered to make inferences on high-dimensional time series. Its non-asymptotic properties are also discussed thoughtfully by balancing the estimation, approximation and truncation errors. Moreover, an alternating gradient descent algorithm with thresholding is designed to search for the high-dimensional estimate, and its theoretical justifications, including statistical and convergence analysis, are also provided. Theoretical and computational properties of the proposed methodology are verified by simulation experiments, and the advantages over existing methods are demonstrated by two empirical examples.

C0358: **Analysis of weighted temporal networks represented by time slices**

Presenter: **Vladimir Batagelj**, IMFM, Slovenia

A problem that often arises in the analysis of weighted networks is the dominance of a few units. It can be avoided by appropriate normalization of network weights (Markov, Salton, Jaccard, Balasza, etc.). Various skeletons (spanning tree, Pathfinder, 1-neighbors, 2-neighbors, etc.) can be used to reveal the basic structure of the network. By analyzing them between successive slices, we can reveal the main changes in the evolution of the temporal network. The approach will be demonstrated by the example of the evolution of the international trade network.

CO101 Room BCB 311 NOVEL PERSPECTIVES IN BAYESIAN STATISTICS Chair: Gavino Puggioni

C0198: **Multivariate isotropic random fields on spheres: Nonparametric Bayesian modeling and L_p fast approximations**

Presenter: **Philip White**, Brigham Young University, United States

Co-authors: Pier Giovanni Bissiri, Emilio Porcu, Galatia Cleanthous, Alfredo Alegria

Multivariate Gaussian random fields defined over d-dimensional spheres are studied. First, we provide a nonparametric Bayesian framework for modeling and inference on matrix-valued covariance functions. We determine the support (under the topology of uniform convergence) of the proposed random matrices, which cover the whole class of matrix-valued geodesically isotropic covariance functions on spheres. We provide a thorough inspection of the properties of the proposed model in terms of (a) first moments, (b) posterior distributions, and (c) Lipschitz continuities. We then provide an approximation method for multivariate fields on the sphere for which measures of accuracy are established. Our findings are supported by simulation studies that show the rate of convergence when truncating a spectral expansion of a multivariate random field at a finite order. To illustrate the modeling framework developed, we consider a bivariate spatial data set of two 2019 NCEP/NCAR Flux Reanalyses.

C0202: **Reliable Bayesian inference in misspecified models**

Presenter: **David Frazier**, Monash University, Australia

A general solution to a fundamental open problem in Bayesian inference is provided, namely poor uncertainty quantification, from a frequency standpoint, of Bayesian methods in misspecified models. While existing solutions are based on explicit Gaussian approximations of the posterior, or computationally onerous post-processing procedures, we demonstrate that correct uncertainty quantification can be achieved by replacing the usual posterior with an intuitive approximate posterior. Critically, our solution is applicable to likelihood-based, and generalised, posteriors as well as cases where the likelihood is intractable and must be estimated. We formally demonstrate the reliable uncertainty quantification of our proposed approach, and show that valid uncertainty quantification is not an asymptotic result and occurs even in small samples. We illustrate this approach through a range of examples, including linear, and generalised, mixed effects models.

C0265: **On the Voigt distribution: Characterization and parameter estimation**

Presenter: **Gavino Puggioni**, University of Rhode Island, United States

Co-authors: Massimo Cannas

The Voigt profile is the convolution of Gaussian and Cauchy random variables. The Voigt is extensively used in atomic and molecular spectroscopy to represent superposition effects. The lack of a moment-generating function and of a closed form for the density has generated some interest in the literature about parameter estimation. We provide a new characterization of the Voigt profile and its associated dual. We also propose an MCMC algorithm to estimate the posterior distribution of both scale and location parameters. A simulation study demonstrates a better performance of our algorithm compared to other approaches.

C0350: **Fast implementation of a general importance sampling algorithm for Bayesian nonparametric models with binary responses**

Presenter: **Dennis Christensen**, University of Oslo, Norway

Co-authors: Per August Moen

Binary response data problems, such as those arising in bioassay, current status data and binary classification, have been an important subfield of Bayesian nonparametrics for the last 50 years. For models based on the Dirichlet process, there exist Markov chain Monte Carlo (MCMC) algorithms given such data. However, for many of the new models developed over the preceding decade, MCMC methods are unavailable when the data comprise both left and right-censored observations. We introduce a new, highly general-importance sampling algorithm which enables

posterior inference for any nonparametric model from which a random sample can be generated. Calculating the importance weights is equivalent to computing the permanents of a class of (0,1)-matrices, which we prove can be done in polynomial time. Furthermore, we provide an efficient implementation of the algorithm by optimising memory management and exploiting sparse data structures. This allows the importance sampling algorithm to handle datasets of size up to several thousand observations.

CO107 Room BCB 309 ADVANCES IN MULTI-VIEW LEARNING AND MIXTURE MODELS Chair: Angela Montanari

C0241: **Finding groups in microbiome data according to multiple data-views**

Presenter: **Silvia Dallari**, Alma mater studiorum- universita di Bologna, Italy

Co-authors: Laura Anderlucci, Angela Montanari

Microbiota plays a crucial role in human health. Next Generation Sequencing technologies have enabled the exploration of the microbiome without isolation and culturing. However, analyzing and translating microbiome data into meaningful biological insights is still challenging due to the data's compositional nature, high dimensionality, sparseness, and over-dispersion. The gut microbiome can vary from individual to individual, and microbiome communities can be grouped to identify community types linked to environmental or health conditions. Different data features, such as individual profiles, community-based descriptors, or genera interactions within a community, provide different perspectives on microbiome complexity. Combining these perspectives may lead to a more comprehensive understanding of microbiome data. The clustering results of the three data views could be combined via consensus clustering or via Bayesian latent structure models. The proposed multi-view clustering method will be applied to a real dataset on the human gut microbiome.

C0288: **Finding the hidden link: Statistical methods for multi-view high-dimensional data**

Presenter: **Katrijn Van Deun**, Tilburg University, Netherlands

Research in many disciplines relies more and more on intensive collections of data representing several points of view. For example, in studying obesity or depression as the outcome of environmental and genetic influences, researchers increasingly collect survey, dietary, biomarker and genetic data from the same individuals. Revealing the variables that are linked throughout these different types of data gives crucial insight into the complex interplay between the multiple factors that determine human behavior, e.g., the concerted action of genes and environment in the emergence of obesity or depression. Although linked high-dimensional multiview data form an extremely rich resource for research, extracting meaningful and integrated information is challenging and not appropriately addressed by current statistical methods. The challenge is to select those variables that are linked throughout the different blocks, and this eludes currently available methods for data analysis. The 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. We will discuss the challenges associated with the analysis of large scale multiview data and present a sparse common, and distinctive components approach to address the challenges.

C0167: **Local moment matching with Gamma mixtures under automatic smoothness penalization**

Presenter: **Oskar Laverny**, Universite Catholique de Louvain, Louvain-la-Neuve, Belgium, Belgium

Co-authors: Philippe Lambert

The class of Erlang mixture has been widely used in the literature for flexible density estimation procedures. More specifically, we consider them for the task of density estimation on the positive real line when the only available information is given as localized moments, such as a histogram with potentially higher-order moments in some bins. By construction, the obtained moment problem is ill-posed and requires regularization. Several penalties can be used for such a task, such as a lasso penalty for the sparsity of the representation, but we focus here on a simplified smoothness penalty coming from the P-splines literature. We show that the corresponding hyperparameter can be selected without cross-validation through the computation of the so-called effective dimension of the estimator, which makes the estimator practical and adapted to these summarized information settings. The flexibility of the local moments representations allows interesting additions, such as enforcing Value-at-Risk and Tail Value-at-Risk constraints on the resulting estimator, making the procedure fit for heavy-tailed estimations.

C0327: **Mixtures of generalised normal distribution with constraints**

Presenter: **Pierdomenico Duttilo**, University G. d'Annunzio of Chieti-Pescara, Italy

Co-authors: Stefano Antonio Gattone, Alfred Kume

A family of univariate mixtures of generalised normal distribution with constrained parameters (CMGND) is proposed. Specifically, the location, scale and shape parameters are constrained to be equal across any subset of mixture components. In this way, it is possible to obtain more parsimonious mixture models and to soften the well-known problem of log-likelihood unboundedness. Additionally, an estimation approach is proposed based on the expectation conditional maximization (ECM) algorithm and the iterative Newton-Raphson method used to handle the non-linear iteration equations of the parameters. A simulation study is performed to assess the estimation performance of a two-component CMGND. Findings show that the estimation accuracy of the constrained mixture is higher than the unconstrained mixture model.

Thursday 24.08.2023 09:00 - 10:00 Parallel Session I – COMPSTAT2023

CC114 Room BCB 307 GENERALIZED LINEAR MODELS Chair: Sara Taskinen

C0345: **Efficient and proper GLM modelling with power link functions**
Presenter: **Vali Asimit**, City University London, United Kingdom
Co-authors: Alexandru Badescu, Feng Zhou

Generalised linear modelling is a flexible predictive model for observational data that is widely used in practice. Such a predictive model requires a careful choice of the link function, and estimation is then achieved by maximum likelihood estimation, for which an optimisation algorithm is required. The computational efficiency is not well-understood, and we raised awareness of the importance of choosing the right link function so the goodness of fit tests and other model adequacy tests are meaningful. The main contributions are as follows: 1) formalise the concept of proper Generalised linear modelling so that a Generalised Linear Model is computationally more reliable, 2) raise awareness of the consequences of choosing an improper link function, 3) provide a novel and efficient numerical algorithm for self-concordant likelihood functions for Poisson and Gamma regression, and 4) provide a novel and efficient numerical algorithm for Inverse-Gaussian regression that violates our definition of properness though the numerical results are computationally stable. The latter two contributions are illustrated through a comprehensive comparison with all available off-the-shelf existing packages implemented in MATLAB, Python and R.

C0362: **Variable importance in generalized linear models: A unifying view using Shapley values**
Presenter: **Christian Kleiber**, Universitaet Basel, Switzerland
Co-authors: Sinan Acemoglu, Joerg Urban

Variable importance in regression analyses is of considerable interest in a variety of fields. There is no unique method for assessing variable importance. However, a substantial share of the available literature employs Shapley values, explicitly or implicitly, for decomposing a suitable goodness-of-fit measure, in the linear regression model, typically the classical R^2. Beyond linear regression, there is no generally accepted goodness-of-fit measure, just a variety of pseudo-R^2s. We formulate and discuss the requirements for goodness-of-fit measures that allow an interpretation of Shapley values in terms of relative and even absolute importance. We suggest employing a pseudo-R^2 based on the Kullback-Leibler divergence, which is of a convenient form for generalized linear models and permits to unify and extend earlier work on variable importance for linear and nonlinear models. We present several examples using data from public health and insurance.

C0378: **Restricted maximum likelihood estimation in generalized linear mixed models**
Presenter: **Luca Maestrini**, The Australian National University, Australia
Co-authors: Francis Hui, Alan Welsh

Restricted maximum likelihood (REML) estimation is a widely accepted and frequently used method for fitting linear mixed models, with its principal advantage being that it produces unbiased estimates of dispersion components. However, the concept of REML does not immediately generalize to the setting of non-normally distributed responses, and it is not always clear the extent to which, either asymptotically or in finite samples, such generalizations reduce the bias of dispersion component estimates compared to standard unrestricted maximum likelihood estimation. We review the various attempts that have been made over the past four decades to develop methods for REML estimation in generalized linear mixed models. We establish four major classes of approaches based on approximate linearization, integrated likelihood, modified profile likelihoods, and direct bias correction of the score function, and show a simulation-based comparison of the approaches.

CC061 Room BCB 308 DESIGN OF EXPERIMENTS Chair: Peter Winker

C0311: **Optimal two-level designs under model uncertainty**
Presenter: **Steven Gilmour**, KCL, United Kingdom
Co-authors: Pi-Wen Tsai

Two-level designs are widely used for screening experiments where the goal is to identify a few active factors which have major effects. We apply the model-robust Q_B criterion for the selection of optimal two-level designs without the requirement of level balance and pairwise orthogonality. We provide a coordinate exchange algorithm for the construction of Q_B-optimal designs for the first-order maximal model and second-order maximal model and demonstrate that different designs will be recommended under different experimenters' prior beliefs. Additionally, we extend the definition of the Q_B criterion to regular and irregular block designs and study the relationship between this new criterion and the aberration-type criteria for blocks. Some trade-offs between orthogonality and confounding will lead to different choice of block designs. Some new classes of model-robust designs which respect experimenters' prior beliefs are found.

C0351: **Nonlinear models for mixture experiments including process variables**
Presenter: **Shroug Alzahrani**, University of Southampton, United Kingdom

Mixture experiments are applied in a variety of fields, for example, food processing, chemical engineering, and product quality improvement, that mix multiple components to perform. The measured response in such experiments is a function not of the amount of the mixture components but their proportions. In some mixture experiments, the blending properties of the mixture may be affected by the processing conditions, such as temperature and pressure, which adds a layer of complexity to the modeling. Such mixture experiments are known as mixture-process variables experiments. For example, while the flavour of a cake depends on the proportions of the cake ingredients, process variables such as cooking time and cooking temperature affect the taste of the cake as well. The experimental region is constrained naturally, as each proportion of mixture components must be greater than zero, and the proportions of all mixture components must sum to one. Often, there are additional restrictions on the proportions when lower and upper limits bound them. A new class of nonlinear models is proposed for mixture-process variables experiments and is compared with standard models from the literature. Moreover, extended forms for modified fractional polynomial models are suggested to fit data from mixture-process variables experiments.

C0390: **Efficient calibration of items in mixed format achievement tests using optimal design methodology**
Presenter: **Ellinor Fackle-Fornius**, Department of Statistics, Sweden
Co-authors: Frank Miller

For large achievement tests, like national tests in school, item calibration is used to determine characteristics of an item, such as difficulty and discrimination. It is important to estimate the item characteristics with as good precision as possible, before the item can be administered in an operational test. We propose an ability-matched item allocation method based on optimal design theory. The method is adapted to handle test items of varying formats analysed using different IRT models, such as the 2-parameter and 3-parameter logistic, as well as graded response models. We demonstrate that the proposed optimal design method leads to increased efficiency and illustrate for which item types it performs best. We also present the results of a real calibration study conducted for the national test in mathematics in Sweden, where the method was evaluated.

CC037 Room BCB 310 BAYESIAN STATISTICS Chair: Eva Cantoni

C0366: **Bayesian inference of sampling weights in COVID-19 testing**
Presenter: **Vasileios Giagos**, University of Essex, United Kingdom

Published COVID-19 testing results provide a daily source of information about the pandemic and, at the initial stages, testing prioritisation has been given to symptomatic patients. This prioritisation introduces an inherent selection preference towards symptomatic cases. We view COVID-19 daily testing results as a weighted sampling process from distinct subpopulations with the sampling weights being the parameters of interest. We incorporate the distribution of the weighted samples, the Wallenius distribution, in a Bayesian setting to estimate and compare their posterior distributions while we address identifiability challenges. In addition, we address computational challenges by proposing an efficient approximation based on the Linear Noise Approximation which demonstrates indistinguishable results under simulated experiments. Finally, we use the Diamond Princess COVID-19 outbreak as a case study to infer the testing priorities of the symptomatic/asymptomatic/healthy groups, and we showcase the flexibility of the Linear Noise Approximation by incorporating the testing mechanism in an epidemiological model to track the dynamics of the COVID-19 outbreak on board the Diamond princess.

C0335: **PCBs intake assessment using a general Bayesian network depending on the meat safety monitoring system**
Presenter: **Hassan Hachem**, Universite Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France, France
Co-authors: Isabelle Albert

Polychlorinated biphenyls (PCBs) are targeted contaminants in the current European monitoring system due to their overall relevance in terms of safety risks, particularly in animal-derived food. As part of SENTINEL, an ANR research project whose objective is to strengthen the monitoring of chemical safety in food, the exposure to PCBs in France associated with the consumption of pork meat is evaluated under the current monitoring system and an alternative one proposed in SENTINEL based on sample pooling. To this end, a general Bayesian network, using discrete and continuous random variables, is proposed to assess the PCBs exposure in pork meat, gathering the sparse contamination and consumption data along the food pathway. The Bayesian modular model considers conventional and organic production meat modes, intake levels, quantities, and frequencies of pork meat consumption. Pseudo-contamination data are introduced in the contamination module to integrate different safety monitoring strategies All models also include historical data already collected as prior information. The results of the models are the dietary consumption exposure which can be compared to evaluate the safety impact of new monitoring strategies. MCMC sampling algorithms were implemented through the R package RJAGS, and pseudo data were produced from R functions. In the future, it will be interesting to fully integrate the monitoring system in the Bayesian model instead of using pseudo data.

C0349: **Federated Bayesian inference for time-to-event data**
Presenter: **Hassan Pazira**, Radboud University Medical Center, Netherlands
Co-authors: Marianne Jonker, Ton Coolen

Due to the limited size of the available survival data sets, especially in rare diseases, it is sometimes challenging to identify the most relevant predictive features using multivariable statistical analysis. This issue may be resolved by combining data from multiple centers into one centralized location without sharing their data with each other, but doing so is difficult in reality because of privacy and security concerns. To address these challenges, we develop and implement a Federated Bayesian Inference (FBI) framework for multi-center data. It aims to leverage the statistical power of larger (combined) data sets without requiring all the data to be aggregated in one location. The FBI framework allows each center to use its own local data to infer the optimal parameter values as well as additional features of the posterior parameter distribution to be able to gather more information which is not captured by alternative techniques. The benefit of FBI over alternative approaches is that, only one inference cycle across the centers is required in FBI. For both simulated and real data, we evaluate how well the suggested technique performs.

CC034 Room BCB 311 COMPUTATIONAL AND FINANCIAL ECONOMETRICS Chair: Massimiliano Caporin

C0151: **Testing beta constancy in capital asset pricing models**
Presenter: **Luis Antonio Arteaga Molina**, Universidad de Cantabria, Spain
Co-authors: Juan Manuel Rodriguez-Poo

A methodology is proposed for testing coefficients constancy in varying coefficient capital asset pricing models with endogenous regressors. The testing procedure is defined as a generalized likelihood ratio that focuses on the comparison of the restricted and unrestricted sum of squared residuals. As a by-product, we have developed a nonparametric method that takes into account the endogenous nature of the regressors to estimate the prices of risk; besides, we establish the asymptotic properties of the estimators. We also investigate the finite sample properties of our test by means of Monte Carlo experiments study and using critical values and p-values estimated using a bootstrap technique. Finally, we apply our test to the Fama and French model using a Fama-French 6 portfolio, sorted by size and book-to-market.

C0360: **Sup-tests against time-varying GARCH models**
Presenter: **Niklas Ahlgren**, Hanken School of Economics, Finland
Co-authors: Alexander Back, Timo Terasvirta

Testing GARCH models against time-varying GARCH models involves nuisance parameters which are not identified under the null hypothesis. Asymptotic distribution theory is used for additive nonlinear regression models to derive misspecification tests against a new GARCH model with a deterministic time-varying intercept. First, we linearise the GARCH model by an ARMA representation. Second, we use testing theory for regression models with additive nonlinearity to derive test statistics. The asymptotic distributions of test statistics can be expressed as functionals of chi-squared processes. The supremum (sup) and average (ave) functionals are used to derive test statistics. The asymptotic distributions of the test statistics are approximated by simulation. In a Monte Carlo study, we find that the proposed sup and ave tests have good size and power properties. The results show that the tests tend to be slightly conservative but have higher power than tests based on auxiliary regressions. The power loss implied by the Taylor expansion in auxiliary regression-based tests is substantial against a time-varying GARCH model with an intercept that is a smooth function of time.

C0153: **Asymptotic inference for new double autoregressive models**
Presenter: **Emma Iglesias**, University of A Coruna (SPAIN), Spain

Extensions of the double autoregressive (DAR) model are proposed. We start with the novel sign-double autoregressive (SDAR) model, in the spirit of the GJR-GARCH model (also named the sign-ARCH model). The new model shares the important property of DAR models where a unit root does not imply nonstationarity and allows for asymmetry. We establish consistency and asymptotic normality of the quasi-maximum likelihood estimator in the context of the SDAR model. Furthermore, it is shown by simulations that the asymptotic properties also apply in finite samples. Finally, an empirical application shows the usefulness of our new model. New DAR models will also be proposed, and the corresponding asymptotic theory and empirical examples will be provided.

CC109 Room BCB 309 TIME SERIES ECONOMETRICS Chair: Davide La Vecchia

C0162: **On a standard method for measuring the natural rate of interest**
Presenter: **Daniel Buncic**, Stockholm University, Sweden

It is shown that Median Unbiased Estimation (MUE), as implemented previously, cannot recover the signal-to-noise ratio parameter of interest and leads to a spurious downward trend in the estimate of the natural rate. We provide a correction to the implementation of MUE in HLW. This correction is quantitatively important and results in substantially smaller point estimates of the signal-to-noise ratio parameter that affects the severity of the downward trend in the natural rate. For the US, the point estimate decreases from 0.040 to 0.013, and is statistically highly

insignificant. For the Euro Area, the UK and Canada, the MUE point estimates are exactly zero. The resulting natural rate estimates from the corrected MUE implementation are up to 100 basis points larger than originally reported.

C0169: **Model averaging prediction for possibly nonstationary autoregressions**

Presenter: **Chu-An Liu**, Academia Sinica, Taiwan

Co-authors: Tzu-Chi Lin

As an alternative to model selection (MS), the focus is on model averaging (MA) for integrated autoregressive processes of infinite order. We derive a uniformly asymptotic expression for the mean squared prediction error (MSPE) of the averaging prediction with fixed weights and then propose a Mallows-type criterion to select the data-driven weights that minimize the MSPE asymptotically. We show that the proposed MA estimator and its variants, Shibata and Akaike MA estimators, are asymptotically optimal in the sense of achieving the lowest possible MSPE. We further demonstrate that MA can provide significant MSPE reduction over MS when the model misspecification bias is algebraic decay. These theoretical findings are supported by Monte Carlo simulations and real data analysis.

C0308: **Semiparametric forecasting using non-Gaussian ARMA models based on s-vines**

Presenter: **Jialing Han**, University of York, United Kingdom

Co-authors: Alexander Alexander John McNeil, Alexandra Dias, Martin Bladt

A semiparametric method for forecasting time series based on the s-vine copula approach for stationary time series developed recently is proposed. By combining a parametric s-vine process to describe serial dependence with a nonparametric model of the marginal distribution, the method offers improved modelling and forecasting for time series that have a non-Gaussian distribution and a nonlinear dependence on past values. The methodology gives a clear meaning to the concept of a non-Gaussian autoregression moving average (ARMA) model in which a parametric object known as the Kendall partial autocorrelation function plays the central role. To demonstrate the potential forecasting gains that can be obtained by using nonGaussian models, an approach to comparing distributional forecasts is applied. The methodology is illustrated with an application to forecasting the force of inflation in the US.

Thursday 24.08.2023 10:30 - 12:30 Parallel Session J – COMPSTAT2023

CO113 Room BCB 307 TUTORIAL II Chair: Francisco Javier Rubio

C0406: **Bayesian variable selection for survival data: Theory, methods, software and applications**
Presenter: **Francisco Javier Rubio**, University College London, United Kingdom

Survival analysis is one of the main branches of Statistics, with applications in medicine, biology, and engineering, to name but a very few. Thanks to recent data linkage and data collection developments, survival data can be enriched with additional individual characteristics, such as sociodemographic, clinical, and genetic information. Thus, it is of interest to select the variables that explain the survival times (i.e. the variables that are associated with the survival times). Two of the most popular models for analysing survival data will be presented: the proportional hazards model and the accelerated failure time model. After that, we will present an overview of different methods for conducting variable selection using these models, including penalised likelihood methods, and Bayesian methods based on spike-and-slab, local, and non-local priors. Then, we will focus the attention on the analysis of the theoretical/asymptotic properties of the methodology based on local and non-local priors, including a discussion of these properties when the model is misspecified (the most common case in practice). Finally, we will illustrate the use of available R packages for conducting variable selection based on these methods and models using simulated and real data. R code will be provided.

CO012 Room BCB 308 NEW TRENDS FOR STATISTICAL COMPUTING: BAYESIAN AND SYMBOLIC DATA ANALYSIS Chair: Yuichi Mori

C0254: **Chestnut plot to visualize aggregated symbolic data**
Presenter: **Junji Nakano**, Chuo University, Japan
Co-authors: Nobuo Shimizu, Yoshikazu Yamamoto

When we have a very large amount of data, we are sometimes interested in comparing meaningful groups of data rather than individual observations. Aggregated symbolic data (ASD) expresses a group of observations that have continuous and categorical variables by using up to the second moments of the variables. The ASD for a group of data is equivalent to the set of means, variances, and correlations for continuous variables, the Burt matrix for categorical variables, and the means of a continuous variable versus a value of a categorical variable. Because ASD with many categorical variables is still complicated, it is preferable to have simple measures of the location and dispersion for a categorical variable, and measures of the correlation between two categorical and/or continuous variables. We propose such measures and use them to visualize ASD using an extension of multiple correspondence analysis. We refer to the proposed graph as a chestnut plot because of the shape of each ASD represented.

C0187: **Monitoring photochemical pollutants based on symbolic interval-valued data analysis**
Presenter: **Liang-Ching Lin**, National Cheng Kung University, Taiwan
Co-authors: Meihui Guo, Sangyeol Lee

The focus is on monitoring photochemical pollutants for anomaly detection based on symbolic interval-valued data analysis. For this task, we construct control charts based on the principal component scores of symbolic interval-valued data. Herein, the symbolic interval-valued data are assumed to follow a normal distribution, and an approximate expectation formula of order statistics from the normal distribution is used in the univariate case to estimate the mean and variance via the method of moments. In addition, we consider the bivariate case wherein we use the maximum likelihood estimator calculated from the likelihood function derived under a bivariate copula. We also establish the procedures for the statistical control chart based on the univariate and bivariate interval-valued variables, and the procedures are potentially extendable to higher dimensional cases. Monte Carlo simulations and real data analysis using photochemical pollutants confirm the validity of the proposed method. The results particularly show the superiority over the conventional method that uses the averages to identify the date on which the abnormal maximum occurred.

C0211: **Mixed-type multivariate Bayesian sparse variable selection with shrinkage priors**
Presenter: **Shao-Hsuan Wang**, National Central University, Taiwan
Co-authors: Hsin-Hsiung Huang, Ray Bai

A Bayesian framework is introduced for mixed-type multivariate regression using shrinkage priors. The proposed method enables joint analysis of mixed continuous and discrete outcomes and facilitates variable selection when the number of covariates p can be much larger than sample size n. Theoretically, we show that the posterior contracts around the true parameter in mixed-response models when p grows subexponentially with n. To cope with the high computational cost when p is large, we introduce a simple two-step variable selection approach. We prove that our two-step algorithm possesses the sure screening property and achieves a faster mixing time than the conventional one-step Gibbs sampler. Moreover, our two-step estimator can provably achieve posterior consistency even when p grows exponentially in n, thus overcoming a limitation of the one-step estimator. Numerical studies and analyses of real datasets demonstrate the ability of our joint modeling approach to improve predictive accuracy and identify significant variables in multivariate mixed response models.

C0212: **Bayesian nonparametric methods for causal effects with intermediate variables**
Presenter: **Chanmin Kim**, SungKyunKwan University, Korea, South

Principal stratification analysis is a method used to estimate causal effects by examining the relationship between treatment and an intermediate variable (such as post-treatment outcomes). However, when the intermediate variable is continuous, parametric modeling methods struggle to capture the complex relationship between the variables. Moreover, separately estimating the outcome and intermediate models leads to uncertainty in the final causal effect estimation. To address these challenges, we propose a fully Bayesian method that uses Bayesian nonparametric models to flexibly estimate all intermediate, outcome, and propensity score models. This method is applicable in both specific and broad confounding situations. We demonstrate the proposed method's performance through a series of simulation studies and apply it to examine the impact of the NOx abatement device (scrubber) installed in US coal-fired power plants on surrounding Ozone concentrations, considering the relationship between the scrubber and NOx emissions from various perspectives.

C0175: **On reliability analysis of one-shot device testing data with defects**
Presenter: **Man Ho Ling**, The Education University of Hong Kong, Hong Kong

The problem of defective devices in the manufacturing industry is considered. Defective devices can arise for various reasons, such as errors made by workers, inadequate quality processes, insufficient training, or issues with reliability during the design stage. The focus is on one-shot device test data that includes defects that have occurred during a realistic manufacturing process. In this scenario, the question arises of whether a failed device is defective or has failed due to its lifetime being shorter than the inspection time. The maximum likelihood approach is explored to estimate the mean-time-to-failure based on a sample of one-shot devices with manufacturing defects. This will be done using gamma and Weibull lifetime distributions. We will also examine how masking can affect the estimation and analysis under different defective rates and masking proportions. In addition, a Bayesian approach is presented to deal with cases with low defective rates.

CO105 Room BCB 310 COMPUTATIONAL ASPECTS OF STRUCTURED MULTIVARIATE AND FUNCTIONAL DATA Chair: Matus Maciak

C0183: **Detection and estimation of changepoints within time-dependent functional profiles**
Presenter: **Matus Maciak**, Charles University, Czech Republic
Co-authors: Sebastiano Vitali

The problem of analyzing financial markets properly relies, among others, on the ability to detect, estimate, and understand different types of structural changes—changepoints. We particularly focus on changes within specific time-dependent functional profiles obtained from the observed options' implied volatility (IV) smiles, and we discuss two different but mutually related approaches: Firstly, consistent statistical tests for detecting significant changepoints are proposed and investigated under different theoretical assumptions and different practical scenarios. Second, an overall stochastic model is postulated, and the unknown (sparse) changepoints are estimated using a regularized quantile estimation framework—all within a model that fully complies with the theory on arbitrage-free markets. Theoretical and empirical aspects are both addressed in detail, and some finite sample performance is illustrated in terms of a simulation study and real data examples.

C0233: **Dependent wild bootstrap for change-point detection in functional time series and random fields**
Presenter: **Martin Wendler**, Otto-von-Guericke University Magdeburg, Germany
Co-authors: Lea Wegner
The aim is to construct a test for the hypothesis of stationarity against the alternative of a location shift in a sequence or fields of dependent, Hilbert-space-valued random variables. We will also consider robust tests, generalizing the Wilcoxon-Mann-Whitney 2-sampe U-statistics to functional data. Since this class of test statistics does not rely on dimension reduction, the limit distribution provides an infinite-dimensional covariance operator as a parameter, which is difficult to estimate. Because of this, we will discuss how the dependent wild bootstrap can be adapted to random fields and to U-statistics with values in a Hilbert space.

C0170: **The minimum weighted covariance determinant estimator for high-dimensional data**
Presenter: **Jan Kalina**, The Czech Academy of Sciences, Institute of Information Theory and Automation, Czech Republic
In a variety of diverse applications, it is very desirable to perform a robust analysis of high-dimensional measurements without being harmed by the presence of a possibly larger percentage of outlying measurements. The minimum weighted covariance determinant (MWCD) estimator, based on implicit weights assigned to individual observations, represents a promising and flexible extension of the popular minimum covariance determinant (MCD) estimator of the expectation and scatter matrix of multivariate data. A regularized version of the MWCD, denoted as the minimum regularized weighted covariance determinant (MRWCD) estimator, is proposed. At the same time, it is accompanied by an outlier detection procedure. The novel MRWCD estimator is able to outperform other available robust estimators in several simulation scenarios, especially in estimating the scatter matrix of contaminated high-dimensional data.

C0221: **Adaptive factor modeling**
Presenter: **Ostap Okhrin**, Technische Universitaet Dresden, Germany
Co-authors: Matthias Fengler
The classical factor model is considered within a sequential change point detection framework that discovers local homogeneity intervals. Our tests for structural breaks in the variance (homogeneity in variance) as well as both in the mean and the variance (complete homogeneity) are based on a maximum statistic of sequential generalized likelihood ratios small-sample distribution of which we approximate by means of a multiplier bootstrap. To handle the high-dimensional parameter problem, we suggest a novel multiplicative bias correction for the multiplier bootstrap. Simulations show that the tests perform very well in terms of size and power. In the empirical application, we study structural breaks for moderately sized equity portfolios.

C0182: **Semi-continuous time series for sparse data with volatility clustering**
Presenter: **Michal Pesta**, Charles University, Czech Republic
Co-authors: Sarka Hudecova
Time series containing a non-negligible portion of possibly dependent zeros, whereas the remaining observations are positive, are considered. They are regarded as GARCH processes consisting of non-negative values. The aim lies in the estimation of the omnibus model parameters taking into account the semi-continuous distribution. The hurdle distribution, together with dependent zeros, causes the classical GARCH estimation techniques to fail. Two different likelihood-based approaches are derived, namely the maximum likelihood estimator and a new quasi-likelihood estimator. Both estimators are proved to be strongly consistent and asymptotically normal. Predictions with bootstrap add-ons are proposed. The empirical properties are illustrated in a simulation study, which demonstrates the computational efficiency of the methods employed. The developed techniques are presented through an actuarial problem concerning sparse insurance claims.

CC111 Room BCB 311 APPLIED ECONOMETRICS Chair: Massimiliano Caporin

C0367: **The asymmetry in the process of price formation: Threshold cointegration analysis**
Presenter: **Emilia Gosinska**, University of Lodz, Poland
Co-authors: Katarzyna Leszkiewicz-Kedzior, Aleksander Welfe
Inflationary processes are sensitive to both supply-side and demand-side shocks as well as instabilities occurring in the markets (e.g. financial crises, pandemics). The coronavirus pandemic resulted in shocks influencing many economies significantly. This must be reflected through adjustments of the parameters of the relevant models, which calls for the appropriate development of the existing methodology. Structural breaks constitute a relatively common problem, especially in empirical studies on transforming economies, that frequently leads to the non-normal distribution of residuals and seriously hinders statistical inference. The main empirical aim is to analyse the determinants of inflation in Poland, allowing for structural breaks and nonlinearities in the long-run relationships. Since most macroeconomic variables are generated by nonstationary stochastic processes and the structural breaks are present in the sample, a nonlinear CVAR model is augmented by adding deterministic terms (representing structural breaks) to the data-generating process. Summing up, the threshold CVAR with deterministic structural breaks is a new tool for describing processes that are currently going on in the global economy. The proper explanation of the inflation processes is a prerequisite to understanding the functioning of economies suffering from exogenous shocks, among which the COVID pandemic seems to be one of the strongest.

C0402: **Revisiting the sources of U.S. imbalances: Wavelet approach**
Presenter: **Jun-Hyung KO**, Aoyama Gakuin University, Japan
The sources of U.S. current account imbalances are investigated using the discrete wavelet method. In line with theoretical implications, the links between the current account and the real interest rate are decomposed into time and frequency domains. The main findings are summarized as follows. First, in the 1980s and the late 2000s, the negative links are strongly observed in more-than-16-year cycles, which supports that the domestic factors are the main source of the U.S. current account deteriorations and improvements. In contrast, from the 1990s to the early 2000s, external factors appear to be the dominant source of current account imbalances. Second, focusing on the domestic factors, the negative trend of the U.S. current account starts with the investment boom and the trough of the current account is accompanied by the saving drought, and finally, the investment drought is attributed to the recovery of the current account imbalances.

C0387: **US equity announcement risk premia**
Presenter: **Lukas Petrasek**, Charles University Prague, Czech Republic
The aim is to analyze the announcement risk premia on the US market. Previous studies have found that a significant portion of the overall risk premia is earned on FOMC meeting days and on days when inflation and employment reports are published. Our evidence suggests that while the announcement risk premia for these days still exists, a much wider range of macroeconomic data releases should be considered. We find that

between September 1987 and March 2023, 99% of the overall cumulative risk premia on the Russell 3000 index is earned on days when data on 17 important macroeconomic variables is released (46% of all trading days). The average return on those days is 6.7 bps compared to 0.9 bps earned on days without any announcements. These results are robust and both economically and statistically significant. We also explore how do the category and frequency of announcements, or the economic cycle influence the results.

C0318: **Good and bad volatility in cryptocurrencies: Connectedness, asymmetry, and their drivers**
Presenter: **Jiri Kukacka**, UTIA AV CR, v.v.i., Czech Republic

Cryptocurrencies exhibit unique statistical and dynamic properties compared to traditional financial assets, making the study of their volatility crucial for portfolio managers and traders. We investigate the volatility connectedness dynamics of a representative set of eight major crypto assets. Methodologically, we decompose the measured volatility into positive and negative components and employ the time-varying parameters vector autoregression (TVP-VAR) framework to be able to show distinct dynamics associated with market booms and downturns. Results suggest that crypto connectedness reflects important events and exhibits more variable and cyclical dynamics than traditional financial markets. Periods of extremely high or low connectedness are clearly linked to specific events in the crypto market and macroeconomic or monetary history. Furthermore, existing asymmetry from good and bad volatility indicates that information about market downturns spills over substantially faster than news about comparable market surges. Overall, the connectedness dynamics is predominantly driven by fundamental crypto factors, while the asymmetry measure also depends on macro factors such as the VIX index and the expected inflation.

C0356: **High frequency financial network connectedness and monetary policy shocks**
Presenter: **Petre Caraiani**, Bucharest University of Economic Studies, Romania

Monetary policy shocks are known to affect financial markets. However, how a monetary policy shock can affect their network structure is less clear. We estimate total daily connectedness and net connectedness for ten industry portfolio indices based on intraday data. Using event-based regressions, we show that total connectedness is positively influenced by surprise changes in interest rates. Net connectedness of some industry indices is also influenced, some in a positive, while others in a negative direction, revealing how monetary policy shocks propagate through the stock market network at a high-frequency level. The results point to the sectoral differences in the propagation of the monetary policy shocks.

CC065 Room BCB 309 ROBUST METHODS Chair: Sara Taskinen

C0273: **Robust monitoring of process dispersion**
Presenter: **Saddam Akber Abbasi**, Qatar University, Qatar
Co-authors: Maha Amouna

Control charts act as the most important tool of the Statistical Process Control (SPC) tool-kit for monitoring process parameters such as location and dispersion. Monitoring process dispersion plays a vital role in improving the quality and production of any manufacturing process. Control chart structures are mostly developed under certain distributional assumptions such as normality and no outliers. When these ideal assumptions are violated, the usual control charts can result in increased false alarms and low detection ability. We will present control chart structures that are based on robust estimation of process dispersion (interquartile range, median absolute deviation, Gini mean, etc.). The performance of these structures is evaluated for a variety of contaminated scenarios and also compared with the usual dispersion chart based on the sample standard deviation. A real-life example will be presented to illustrate the working of new control chart structures. The results will be helpful for quality control practitioners to choose a robust control chart for monitoring process dispersion.

C0309: **Distributionally robust halfspace depth**
Presenter: **Pavlo Mozharovskyi**, LTCI, Telecom Paris, Institut Polytechnique de Paris, France
Co-authors: Jevgenijs Ivanovs

Statistical data depth function measures the centrality of an observation with respect to a distribution or a data set by a number between 0 and 1 while satisfying certain postulates regarding invariance, monotonicity, and convexity. It constitutes a contemporary domain of rapid development to meet growing demand in various areas of industry, economy, social sciences, etc. Being one of the most studied depth notions, Tukey's halfspace depth can be seen as a stochastic program, and as such, it suffers from the optimizer's curse, so that a limited training sample may easily result in a poor out-of-sample performance. We propose a generalized halfspace depth concept relying on the recent advances in distributionally robust optimization, where every halfspace is examined using the respective worst-case distribution in the Wasserstein ball centered at the empirical law. This new depth can be seen as a smoothed and regularized classical halfspace depth, which is retrieved as the radius of the Wasserstein ball vanishes. It inherits the main properties of the latter and, additionally, enjoys various new attractive features such as continuity and strict positivity beyond the convex hull of the support. We provide numerical illustrations of the new depth and its advantages and develop some fundamental theory. In particular, we study the upper-level sets and the median region, including their breakdown properties.

C0289: **A robust combined nonparametric method for comparing two locations**
Presenter: **Marco Marozzi**, Ca' Foscari University of Venice (Italy), Italy

Traditional methods for comparing two samples are based on sample means, which are very non-robust estimators for population means. Non-normality and different variances adversely affect the size and power of traditional tests. Trimmed means are robust and have high efficiency relative to population means. The Yuen test is a familiar method based on trimmed means. There is no agreement about the preferable trimming rate. The power of several Yuen tests with different trimming rates is studied. In general, there is no test being uniformly best. Different tests are sensitive to different features of the data. Therefore, no test can be expected to perform well in all possible situations. In fact, the various Yuen tests show very different power for different distributions because the best trimming rate depends on distribution tail weight. In this context, combined testing is appealing because it aims at providing a test, based on the combination of several tests -each with at least one good feature- that inherits the good features of the single tests being combined. The combined test is expected to perform well in many different situations, with a power that is always larger than the least powerful single test and very often similar to the most powerful single test. Therefore, a bootstrap test based on the combination of several Yuen tests is presented. It is shown that the combined test is powerful irrespective to the underlying distribution.

C0184: **Robustness under missing data: A comparison with special attention to inference**
Presenter: **Carole Baum**, ULiege, Belgium
Co-authors: Arnout Van Messem, Holger Cevallos-Valdiviezo

Missing value imputation is a highly studied topic. A plethora of techniques have been proposed over the years to find suitable values to replace missing data. Nowadays, imputation techniques are widely used, but a large-scale comparison of these methods - especially in terms of their robustness against outliers - seems to be missing. During a first attempt to fill this gap, we evaluate a large selection of imputation techniques involving classic and robust procedures by means of a simulation study with continuous data and different configurations of missing data and outliers. To evaluate the imputation capability and robustness of the imputation techniques, we computed the error between the original and the imputed values. However, often, the main concern is on the analysis that is performed after imputation. Therefore, in the second phase of our research, we evaluated the inferences and predictions made by different robust regression methods combined with an imputation technique in a simulation study. Both row-wise and cellwise outliers were generated, so we considered in the evaluation row-wise robust regression techniques as well as cellwise robust regression techniques. To evaluate the combined regression and imputation strategies in terms of inference capability, we measured the bias and variance of the estimated regression coefficients.

C0339: **Estimation of treatment effects based on robust sparse reduced-rank regression**
Presenter: **Ryoma Hieda**, Doshisha University, Japan
Co-authors: Shintaro Yuki, Kensuke Tanioka, Hiroshi Yadohisa
In clinical trials, we are interested in the estimation of heterogeneous treatment effects (HTE) to develop strategies for personalized medicine. We focus on the modified covariate method (MCM) as one of the methods to estimate the HTE. The model of MCM includes the term of interaction between the treatment and covariates without the main effects and is formulated for a single outcome. However, in clinical trials, there can be interest in multiple outcomes, such as primary and secondary endpoints. Therefore, we extended MCM to the case of multiple outcomes. In addition, we observed that data from clinical trials could include outliers and highly correlated outcomes. In such cases, the HTE cannot be properly estimated. Hence, we propose a method to estimate HTE using MCM in a framework of Robust sparse reduced-rank regression. The proposed method improves the accuracy of estimating HTE because it can deal with highly correlated outcomes by setting rank constraints on the regression coefficient matrix for treatment effects and removing the effects of outliers. We demonstrate the effectiveness of the proposed method based on simulation and real data examples.

CP001 Room Poster session POSTER SESSION Chair: Cristian Gatu

C0214: **BayMDS: An R package for Bayesian multidimensional scaling and choice of dimension**
Presenter: **Man-Suk Oh**, Ewha Womans University, Korea, South
Over the last two decades, there has been a great interest in Bayesian approaches to multidimensional scaling (MDS) due to their advantages over traditional MDS methods. It provides an object configuration along with estimation errors and a simple Bayesian dimension selection criterion MDSIC for optimal dimensionality. However, Bayesian MDS (BMDS) requires a complicated Markov chain Monte Carlo (MCMC) method that may prohibit the wide use of BMDS by practitioners. A set of R functions to perform BMDS, using WinBUGS for MCMC is available. However, WinBUGS has not been updated since 2007 and it may not be efficient since it does not consider special characteristics of BMDS model. In view of these considerations, we have developed an R package bayMDS to implement BMDS that is efficient and can be easily applied by non-experts in MCMC.

C0267: **Modelling symbiotic species richness from invertebrate aquatic hosts using generalized linear and additive models**
Presenter: **Svitlana Shvydka**, Slovak University of Technology in Bratislava, Slovakia
Co-authors: Volodimir Sarabeev, Mykola Ovcharenko, Maria Zdimalova
The data on parasites and epibiotic organisms from the gammarid (Amphipoda) hosts were analyzed by the GLM and GAM. In order to understand the dynamics of symbiotic communities in the host, the species richness was analysed at the individual level by applying Poisson distributions. As explanatory covariates, host characteristics and density, water parameters, geographical locality and habitat type were used. The best models showed that the richness of the whole symbiotic community was positively associated with host body length and concentration of nitrates in water (0.03-1.6 mg/dm3), but negatively with temperature (1.5-15C), pH (7.1-8.4) and total phosphorus (0.04-0.23 mg/dm3). Host species, habitat and locality also had a significant effect on the symbiotic species richness. There was no significant association of the species richness with water parameters (conductivity, concentration of chlorophylla, nitrogen, organic nitrogen, ammonium nitrogen, organic phosphorus, mineral phosphorus) and host density. The modelling approach using both types of models, the GLM and GAM, offers a reliable tool for understanding the effect of biotic and abiotic factors on the species richness of symbiotic organisms. S.S. and V.S. are funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under project Nos. 09I03-03-V01- 00029 and 09I03-03-V01-00017, respectively.

C0332: **watson: An R package for fitting mixtures of Watson distributions**
Presenter: **Lukas Sablica**, WU Vienna University of Economics and Business, Austria
Co-authors: Kurt Hornik, Josef Leydold
The purpose is to present and showcase the R package "watson", which provides a computational framework for fitting and random sampling of the Watson distribution on a p-dimensional sphere. We first introduce the random sampling scheme of the package, which offers two sampling algorithms. What is more, the package offers a smart tool to combine these two methods, and based on the selected parameters, it approximates the relative sampling speed for both methods and picks the faster one. In addition, we describe the main fitting function for the mixtures of Watson distribution, which uses the expectation-maximization (EM) algorithm. Special features are the possibility to use multiple variants of the E-step and M-step, sparse matrices for the data representation and a control parameter which will dynamically eliminate small clusters with overall contribution smaller than this parameter. Moreover, we discuss the numerical issues of the whole fitting procedure and describe how this is handled and solved in the package. Finally, we demonstrate the package on multiple examples involving misspecified simulation study, estimation of the New Zealand earthquake data and depth image clustering.

C0385: **Comparing F1-scores of more than two binary medical tests**
Presenter: **Kanae Takahashi**, Hyogo Medical University, Japan
In the medical field, binary classification problems are common, and accuracy, sensitivity, specificity, and negative and positive predictive value are often used as indicators of the performance of binary medical tests. Additionally, the F1-score is often used in the field of information retrieval, and it is gaining popularity in the medical field. This score is defined as the harmonic mean of recall (sensitivity) and precision (positive predictive value). A statistical test procedure to compare two F1-scores was recently proposed. However, it is often the case that more than two F1-scores are reported and considered at the same time, and it may be desirable to compare them simultaneously. Therefore, using the multivariate central limit theorem and the delta-method, we developed a test procedure for comparing F1-scores of more than two binary medical tests simultaneously.

C0323: **Optimal consumption and investment with independent stochastic labor income**
Presenter: **Seyoung Park**, University of Nottingham, United Kingdom
Co-authors: Alain Bensoussan
The aim is to develop a new dynamic continuous-time model of optimal consumption and investment to include independent stochastic labor income. We reduce the problem of solving the Bellman equation to the problem of solving an integral equation. We then explicitly characterize the optimal consumption and investment strategy as a function of the income-to-wealth ratio. We provide some analytical comparative statics associated with the value function and optimal strategies. We also develop a quite general numerical algorithm for control iteration and solve the Bellman equation as a sequence of solutions to ordinary differential equations. This numerical algorithm can be readily applied to many other optimal consumption and investment problems, especially with extra nondiversifiable Brownian risks, resulting in nonlinear Bellman equations. Finally, our numerical analysis illustrates how the presence of stochastic labor income affects the optimal consumption and investment strategy.

Thursday 24.08.2023 14:15 - 15:45 Parallel Session K – COMPSTAT2023

CV032 Room BCB 311 MACHINE LEARNING AND COMPUTATIONAL METHODS Chair: Rosaria Lombardo

C0373: **Goodness-of-fit and clustering of spherical and directional data: A comprehensive R package**
Presenter: **Giovanni Saraceno**, University at Buffalo, United States
Co-authors: Marianthi Markatou

A new R package is presented that encodes innovative methodologies for data analysis. The package offers a comprehensive implementation of goodness-of-fit tests and clustering techniques based on quadratic distances. One-sample tests and two-sample tests for assessing the fit of probability distributions are implemented. Furthermore, tests for uniformity on the d-dimensional sphere based on Poisson kernel densities, provide additional capabilities. The package incorporates a clustering algorithm designed for data that can be analyzed as spherical data. By leveraging a mixture of Poisson-kernel-based densities on the sphere, the method facilitates effective clustering of spherical (or spherically transformed) data, providing insights into the underlying patterns and relationships. In summary, the proposed R package encompasses a suite of tools through which researchers and practitioners can gain deeper insights, make robust inferences, and provide potentially impactful analyses across diverse fields.

C0285: **Fairness in machine learning in the presence of missing values**
Presenter: **Aeysha Bhatti**, University of Stellenbosch, South Africa

The fairness of Machine Learning algorithms is a topic that is receiving increasing attention, as more and more algorithms permeate the day-to-day aspects of our lives. One way in which bias can manifest in a data source is through missing values. If data are missing, these data are often assumed to be missing completely randomly, but usually, this is not the case. In reality, the propensity of data being missing is often tied to socio-economic status or demographic characteristics of individuals. There is very limited research into how missing values and missing value handling methods can impact the fairness of an algorithm. We conduct a systematic study starting from the foundational questions of how the data are missing, how the missing data are dealt with and how this impacts fairness, based on the outcome of a few different types of machine learning algorithms. Most researchers, when dealing with missing data, either apply listwise deletion or tend to use the simpler methods of imputation versus the more complex ones. We study the impact of these simpler methods on the fairness of algorithms. We also investigate the impact of these methods on the trade-off between accuracy and fairness.

C0306: **A novel approach for estimating functions in the multivariate setting based on an adaptive knot selection for B-splines**
Presenter: **Mary Savino**, University Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay and Andra, France
Co-authors: Celine Levy-Leduc

A novel data-driven method is outlined for estimating functions in a multivariate nonparametric regression model based on an adaptive knot selection for B-splines. The underlying idea of our approach for selecting knots is to apply the Generalized Lasso, since the knots of the B-spline basis can be seen as changes in the derivatives of the function to be estimated. This method was then extended to functions depending on several variables by processing each dimension independently, thus reducing the problem to a univariate setting. The regularization parameters were chosen by means of a criterion based on EBIC. The nonparametric estimator was obtained using a multivariate B-spline regression with the corresponding selected knots. Our procedure was validated through numerical experiments by varying the number of observations and the level of noise to investigate its robustness. The influence of observation sampling was also assessed, and our method was applied to a chemical system commonly used in geoscience. For each different framework considered in this presentation, our approach performed better than state-of-the-art methods. Our completely data-driven method is implemented in the GLOBER R package, which will soon be available on the Comprehensive R Archive Network (CRAN).

C0363: **Computer algebra systems in R**
Presenter: **Soeren Hoejsgaard**, Aalborg University, Denmark
Co-authors: Mikkel Meyer Andersen

R's ability to do symbolic mathematics is largely restricted to finding derivatives. There are many tasks involving symbolic math that are of interest to R users, e.g. inversion of symbolic matrices, limits and solving non-linear equations. Users must resort to other computer algebra systems (CAS) for such tasks, and many R users(especially outside of academia) do not readily have access to such software. There are also other indirect use cases of symbolic mathematics in R that can exploit other strengths of R, including Shiny apps with auto-generated mathematics exercises. We maintain two R-packages that enable symbolic mathematics in R: Ryacas and caracas. Ryacas is based on Yacas (Yet Another Computer Algebra System), and caracas is based on SymPy (Python library). Each has its advantages: Yacas is extensible and has a close integration to R, which makes auto-generated mathematics exercises easy to make.SymPy is feature-rich and thus gives many possibilities. We will discuss the two packages and demonstrate various use-cases, including uses that help to understand statistical models of Shiny apps with auto-generated mathematics exercises.

CI002 Room BCB 307 ROBUST STATISTICS FOR MODERN INFERENCE PROBLEMS Chair: Eva Cantoni

C0168: **The influence function of graphical lasso estimators**
Presenter: **Ines Wilms**, Maastricht University, Netherlands
Co-authors: Gaetan Louvet, Jakob Raymaekers, Germain Van Bever

The precision matrix that encodes conditional linear dependency relations among a set of variables forms an important object of interest in multivariate analysis. Sparse estimation procedures for precision matrices, such as the graphical lasso (Glasso), gained popularity as they facilitate interpretability, thereby separating pairs of variables that are conditionally dependent from those that are independent (given all other variables). Glasso lacks, however, robustness to outliers. To overcome this problem, one typically applies a robust plug-in procedure where the Glasso is computed from a robust covariance estimate instead of the sample covariance, thereby providing protection against outliers. In this talk, we study such estimators theoretically by deriving and comparing their influence function, sensitivity curves and asymptotic variances.

C0190: **Some aspects of robust optimal transportation, with applications to statistics and machine learning**
Presenter: **Davide La Vecchia**, University of Geneva, Switzerland

Optimal transport (OT) theory and the related p-Wasserstein (Wp) distance are popular tools in statistics and machine learning. Recent studies have been remarking that inference based on OT and on Wpis sensitive to outliers. To cope with this issue, we work on a robust version of the primal OT problem (ROBOT) and show that it defines a robust version of W1, called robust Wasserstein distance, which is able to down-weight the impact of outliers. We study the properties of this novel distance and use it to define minimum distance estimators. Our novel estimators do not impose any moment restrictions: this allows us to extend the use of OT methods to inference on heavy-tailed distributions. We also provide statistical guarantees of the proposed estimators. Moreover, we derive the dual form of the ROBOT and illustrate its applicability to machine learning. Numerical exercises provide evidence of the benefits yielded by our methods.

C0280: **Robust cellwise regularized sparse regression**
Presenter: **Samuel Muller**, Macquarie University, Australia
Co-authors: Peng Su, Tarr Garth, Suojin Wang

The robust variable selection currently has some focus on dealing with cellwise contamination in the design matrix where only some but not all

elements of an observation vector are contaminated. The problem is particularly challenging when the number of variables is large. Traditional robust methods can fail when the data is high-dimensional and too many observation rows experience some cellwise contamination. We explore how using initial robust empirical covariance matrix estimators, together with regularization approaches, helps in robustly selecting variables by simultaneously shrinking regression coefficients and identifying outlying cells in the data matrix. Specifically, we highlight the performance of CR-Lasso, a new approach which incorporates a constraint on the deviation of each cell in the loss function to detect outliers based on regression residuals and cell deviations by combining L1 and cellwise outlier regularization.

CO013 Room BCB 308 NEW DEVELOPMENTS IN BAYESIAN ANALYSIS Chair: Ray-Bing Chen

C0154: **Efficient data augmentation techniques for some classes of state space models**

Presenter: **Siew Li Linda Tan**, National University of Singapore, Singapore

Data augmentation improves the convergence of iterative algorithms, such as the EM algorithm and Gibbs sampler, by introducing carefully designed latent variables. We first propose a data augmentation scheme for the first-order autoregression plus noise model, where optimal values of working parameters introduced for recentering and rescaling of the latent states, can be derived analytically by minimizing the fraction of missing information in the EM algorithm. The proposed data augmentation scheme is then utilized to design efficient Markov chain Monte Carlo (MCMC) algorithms for Bayesian inference of some non-Gaussian and nonlinear state space models, via a mixture of normals approximation coupled with a block-specific reparametrization strategy. Applications on simulated and benchmark real datasets indicate that the proposed MCMC sampler can yield improvements in simulation efficiency compared with centering, noncentering and even the ancillarity-sufficiency interweaving strategy.

C0194: **A modified VAR-GARCH model for asynchronous multivariate financial time series via variational Bayesian inference**

Presenter: **Lai Wei-Ting**, National Central University, Taiwan

A modified VAR-GARCH model, called M-VAR-GARCH, is proposed for modeling asynchronous multivariate financial time series with GARCH effects and simultaneously accommodating the latest market information. A variational Bayesian (VB) procedure is developed to infer the M-VAR-GARCH model for structure selection and parameter estimation. We conduct extensive simulations and empirical studies to evaluate the fitting and forecasting performances of the M-VAR-GARCH model. The simulation results reveal that the proposed VB procedure produces satisfactory selection performances. In addition, our empirical studies find that the latest market information in Asia can provide helpful information to predict market trends in Europe and South Africa, especially when momentous events occur.

C0203: **Graphical copula GARCH modeling with dynamic conditional dependence**

Presenter: **Mike So**, The Hong Kong University of Science and Technology, Hong Kong

The aim is to develop a graphical copula GARCH model for volatility modeling. To allow high-dimensional modeling for large portfolios, the complexity of the modeling is greatly reduced by introducing conditional independence among stocks given the market risk factors, such as the S&P500 index in the United States. The market risk factors are modeled using a directed acyclic graph (DAG) model with a pairwise-copula construction to allow flexible distributional modeling. The use of the DAG model gives a topological order to the market risk factors, which can be regarded as a list of directions of the flow of information or disturbance. The conditional distributions among stock returns are also modeled through pairwise-copula constructions for flexibility. We adopt dynamic conditional dependence structures to allow the parameters in the copulas to be time-varying such that we can model dynamically the tail dependence between any two stocks. Three-stage estimation is used for estimating parameters in the marginal distributions, the copulas of the DAG of the market risk factors, and the copulas of the stocks. Bayesian inference is used to learn the structure of the DAG. In the simulation study, we show that these estimation procedures can be used to recover the parameters and the DAG accurately. With Bayesian inference, we can allow the structure of the market risk factors to be random, and model averaging can be done to obtain robust predictions of volatility.

C0228: **Bayesian analysis of multivariate longitudinal binary data**

Presenter: **Kuo-Jung Lee**, National Cheng Kung University, Taiwan

A Bayesian multivariate probit linear mixed model is proposed to analyze multivariate longitudinal binary data. We estimate the effects of the covariates on the responses while accounting for three types of complex correlations present in the data. These include the correlations within separate responses over time, cross-correlations between different responses at different times, and correlations between different responses at each time point. The correlation matrix is estimated using hypersphere decomposition to meet the positive definiteness constraint. Simulations and real examples are used to demonstrate the proposed methods.

CO106 Room BCB 310 ADVANCES IN FUNCTIONAL DATA: THEORY AND APPLICATIONS Chair: Enea Bongiorno

C0312: **On specifying a link function of a single functional index model**

Presenter: **Kwo Lik Lax Chan**, Universita degli Studi del Piemonte Orientale, Italy

An important task in regression analysis is to choose the right specification of the link function that models the dependence structure between the random elements. A challenge arises in the framework of scalar on function regression as the link function is a real-valued operator acting on a functional space, and it is difficult to visualize and hence select a coherent specification. A specification test that uses a semi-parametric approach is proposed, in particular by exploiting the Single Functional Index Model. The test statistic is a special form of U-statistic; its asymptotic null distribution is derived under suitable assumptions, and consistency is proved for a specific class of alternatives. The finite sample performances of the test are evaluated through a simulation study using both asymptotic p-values and some bootstrap approaches. An application of the method developed to a problem commonly arises in the food industry is performed to demonstrate the potentialities of the method.

C0209: **Orthogonal decomposition of multivariate densities in Bayes spaces in context of functional data analysis**

Presenter: **Karel Hron**, Palacky University, Czech Republic

Co-authors: Christian Genest, Johanna Neslehova

Probability density functions can be embedded in the geometric framework of Bayes spaces which respect their relative nature and enable further modeling and analysis. Specifically, the Hilbert space structure of Bayes spaces has several important implications for estimation theory, Bayesian statistics as well as functional data analysis. In this contribution, an orthogonal decomposition of multivariate densities in Bayes spaces using a distributional analog of the HoeffdingSobol identity is constructed. The decomposition is based on reformulation of the standard (arithmetic) marginals to so-called geometric marginals, which are orthogonal projections of the univariate information contained in multivariate densities, follow the Yule perturbation and coincide with the arithmetic ones in case of independence. Accordingly, the decomposition contains an independent part and all possible interaction terms. The orthogonality of the decomposition results in Pythagoras' Theorem for squared norms of the decomposed densities and margin-free property of the interaction terms. There is also a relation between copula-based representation of densities and their functional data analysis. The latter will be illustrated with empirical geochemical data.

C0222: **Combining concurrent and functional linear regression**

Presenter: **Sven Otto**, University of Cologne, Germany

Co-authors: Alois Kneip, Dominik Liebl

A new function-on-function linear regression model that incorporates common and point effects of a regressor function on a response function is introduced. The model comprises two components: a Hilbert-Schmidt integral operator for the common component and a concurrent component

that captures the regressor's impact on the response at each domain point. These components are separately identifiable under mild conditions. We propose a smoothing spline estimator, provide asymptotic theory, and demonstrate the model's practicality using sports data.

C0245: **Additive regression with general imperfect variables**
Presenter: **Germain Van Bever**, Universite de Namur, Belgium
Co-authors: Jeong Min Jeon

An additive model is presented where the response variable is Hilbert-space-valued, and predictors are multivariate Euclidean, and both are possibly imperfectly observed. Considering Hilbert-space-valued responses allows us to cover Euclidean, compositional, functional and density-valued variables. By treating imperfect responses, we can cover functional variables taking values in a Riemannian manifold and the case where only a random sample from a density-valued response is available. This treatment can also be applied in semiparametric regression. Dealing with imperfect predictors allows us to cover various principal component and singular component scores obtained from Hilbert-space-valued variables. For the estimation of the additive model having such variables, we use the smooth backfitting method. We provide full non-asymptotic and asymptotic properties of our regression estimator and present its wide applications via several simulation studies and real data applications.

CC050 Room BCB 309 FORECASTING Chair: Nicolas Hernandez

C0374: **Skill scores, predictive power and limits of predictability**
Presenter: **Tanja Zahn**, Goethe University Frankfurt, Germany
Co-authors: Marc-Oliver Pohle

Forecasts are usually issued over multiple forecast horizons. Often, they range quite far into the future, giving rise to the questions of how useful long-horizon forecasts are and up to which maximum horizon forecasting is really sensible. A methodology is developed to answer this question for arbitrary types of forecasts, including univariate and multivariate mean, quantile and probabilistic forecasts. The key tool is a nicely interpretable measure of predictive power, which amounts to the ratio of variation explained by the forecasts to overall variation in the variable of interest. This measure nicely complements the standard toolkit of forecast evaluation, informing about the usefulness of the forecasts in the first place. Measures for the information content of forecasts and predictability of the variable of interest are introduced, and their relationship to predictive power is discussed. The methodology is applied to macroeconomic forecasts. The analysis shows that the predictive power of state-of-the-art forecasting methods for inflation and GDP growth can be very limited even for short forecast horizons, which hints at a lack of predictability of the economy.

C0392: **A strong link between mortality projections and frailty in Lee Carter model**
Presenter: **Maria Carannante**, University of Salerno, Italy
Co-authors: Valeria D Amato, Steven Haberman, Massimiliano Menzietti

Discrepancies due to the misspecification of the mortality model are known as model risk. Model risk includes shocks caused by period effects that temporarily change the mortality behaviours. Period effects could fit the definition of frailty, according to which is the set of unobservable factors that determines the heterogeneity in mortality. In actuarial literature, frailty is an unobserved variable defined by a non-negative random variable. We propose the use of the Age-specific and Temporal Frailty Lee Carter (ATFLCA) model to estimate the differences in mortality forecasts between the general population and the annuity population. The underlying idea is that the use of frailty parameters in LC models, identified as the co-morbidity status of the individuals, can detect the differences in terms of adverse selection in the life insurance market since people with a lower risk of death are more likely to own life insurance.

C0396: **Nowcasting GDP in Switzerland: What are the gains from machine learning algorithms?**
Presenter: **Rolf Scheufele**, Swiss National Bank, Switzerland
Co-authors: Milen Arro

Several machine learning methods for nowcasting GDP in Switzerland are compared. Based on a large mixed-frequency data set, we investigate the predictive ability of regression-based methods (Ridge, LASSO, Elastic net), tree-based methods, bagging and SVR. As benchmarks, we use univariate models, forward selection algorithms and factor models. For the period between the Financial Crisis and the COVID-19 crisis, which is particularly challenging in terms of nowcasting, we find that all considered ML techniques beat the univariate benchmark and the forward selection algorithms. Ride regression and SVR turn out to be most successful and outperform the factor model based on principle components by more than 10% in terms of RMSE.

C0394: **Multivariate forecast for financial stock prices: A hybrid VAR-LSTM deep learning model**
Presenter: **Diana Mendes**, ISCTE-IUL, Portugal
Co-authors: Vivaldo Mendes, Nuno Ferreira

The forecasting of stock price dynamics is a challenging task since these kinds of financial datasets are characterized by irregular fluctuations, nonlinear patterns, and high uncertainty changes. Deep neural network models, particularly the LSTM (Long Short Term Memory) algorithm, have been increasingly used by researchers and practitioners to analyze, trade, and predict financial time series, defining a new essential tool in several sectors' decision-making processes. The primary purpose focuses on a multivariate forecast of the U.S. stock index S&P500, using Nasdaq, Dow Jones, and U.S. treasury bills for three months yields of the secondary market series, with daily frequency, between January 2018 and April 2023. With the support of a hybrid windowed VAR (Vector Auto Regressive) trend corrected by an LSTM recurrent neural network, we consistently obtain low forecast errors (around 4%), even during the COVID-19 crisis. In addition, nonlinear Granger causality, based on transfer entropy, was tested between the periods with strong intervention by the Federal Bank, concluding that yield variation Granger causes the stock indices returns. In contrast, this causal relationship outside these periods was inverted, with the indices' returns causing yield variation.

Thursday 24.08.2023 16:15 - 17:45 Parallel Session L – COMPSTAT2023

CV044 Room Virtual room R01 APPLIED STATISTICS AND ECONOMETRICS Chair: Rosaria Lombardo

C0357: **The transitory component of health care employment**
Presenter: **Luiggi Donayre**, University of Minnesota - Duluth, United States
Co-authors: Lacey Loomer

U.S. healthcare employment is decomposed into permanent and transitory components to study the short-run properties of the deviation of healthcare employment from its long-run trend and to understand its correlation with the aggregate business cycle for the 1990-2023 period. To do so, we consider a battery of linear and nonlinear time series models that appropriately address the outlier observations associated with the COVID-19 pandemic. Based on this model-based approach, we find empirical support for the presence of asymmetries in the behavior of the transitory component of healthcare employment. Because several alternative specifications produce a similar fit, but generate measures with different shapes and depths, we construct a model-average measure to address model-based uncertainty. This estimated transitory component exhibits a moderately positive correlation with measures of the aggregate business cycle, suggesting that healthcare employment responds to overall economic conditions only partially. Furthermore, this correlation varies over time, indicating that recessions are not all alike in their effect on the healthcare sector.

C0381: **Financial distress prediction using machine learning: When Altman meets Merton in a transition economy**
Presenter: **Minh Nguyen**, University of Hawaii at Manoa, United States

The purpose is to explore financial distress prediction for public firms in Vietnam as a transition economy by combining variables from the Altman and Merton model. In order to do that, we employ four machine learning methods, including linear discriminant analysis, logistic regression, support vector machines, and neural networks. Our results show that in most cases, the models combining the Altman's and Merton's variables outperform those that only use either of these lists of variables.

C0283: **Data-driven optimal phase division for improved weather index insurance design**
Presenter: **Jing Zou**, Technische Universitaet Dresden, Germany
Co-authors: Ostap Okhrin, Martin Odening

Past research has demonstrated that the hedging effectiveness of weather index insurance can be improved by decomposing the vegetation period of plants into separate growth phases. We conduct a data-driven selection of optimal division points to improve further the division procedure to mitigate temporal basis risk. Various statistical and machine learning methods are applied and compared concerning their ability to model the weather-yield relation. Using farm-level winter barley yield data from 217 rural households in Saxony, Germany, and corresponding weather data and phenology information, we first separate the whole crop growth cycle of winter barley into four sub-phases according to the phenology reports, secondly, fix the start and end points while relax the internal three points to create 24804 combinations, thirdly search for the optimal division points by employing Polynomial Regression, Generalized Additive Model, Random Forests, Support Vector Machine, and Artificial Neural Networks to model the nonlinear relationship between 8 weather variables and yield variability. The results suggest that optimal division models achieve better performance than benchmark models, and the consistency in data-driven flexible points with the reference ones, especially in terms of Polynomial Regression, Generalized Additive Models and Artificial Neural Networks. In addition, the model fitting results of different methods indicate robustness.

C0313: **Studying the COVID-19 lockdown effects on Iranian traffic behavior in three calendars with functional data analysis**
Presenter: **Mohammad Fayaz**, Allameh Tabataba'i University, Iran

The COVID-19 lockdown has affected many aspects of people's lives, like air pollution, economics, traffic, etc. We have collected spatio-temporal traffic datasets between provinces in Iran from March 2010 to January 2023 from more than 2500 count data stations. We have studied four indices, including total traffic, and the number of traffic offenses - speeding, unsafe distance, and Illegal overtaking in the following time periods: before, during, and after COVID-19 lockdowns. The Iranian official calendar is the Solar Hijri calendar, while the Islamic Hijri calendar and Gregorian calendar are also important because many holidays and events are occurred due to them. The time series decomposition methods that consider multiple seasonal periods are applied. The outlier detection methods from fda.usc R package is used. The forecasting scenario is without the COVID-19 lockdowns and estimates the lockdown effects. The Interval-wise testing (IWT) results for comparing before and after lockdown are presented with adjusted p-values. The spatial variability of count stations near roads is estimated with functional and spatial statistical methods. The R codes are available on the author's GitHub page for reproducible purposes.

CI005 Room BCB 307 BAYESIAN MODELS: INFERENCE AND APPLICATIONS Chair: Ioanna Manolopoulou

C0348: **A Bayesian non-parametric approach for causal mediation with a post-treatment confounder**
Presenter: **Michael Daniels**, University of Florida, United States
Co-authors: Woojung Bae

A new Bayesian non-parametric (BNP) method is proposed for estimating the causal effects of mediation in the presence of a post-treatment confounder. We specify an enriched Dirichlet process mixture (EDPM) to model the joint distribution of the observed data (outcome, mediator, post-treatment confounder, treatment, and baseline confounders). For identifiability, we use the extended version of the standard sequential ignorability. The observed data model and causal identification assumptions enable us to estimate and identify the causal effects of mediation, i.e., the natural direct effects (NDE), and indirect effects (NIE). Our method enables easy computation of NDE and NIE for a subset of confounding variables and addresses missing data through data augmentation under the assumption of ignorable missingness. We conduct simulation studies to assess the performance of our proposed method. Furthermore, we apply this approach to evaluate the causal mediation effect in the Rural LITE trial, demonstrating its practical utility in real-world scenarios.

C0397: **Bayesian regression tree ensembles for survival analysis**
Presenter: **Richard Hahn**, Arizona State University, United States

Aspects of time-to-event data that make it challenging for non-parametric covariate adjustments are reviewed. Recent work using tree-ensembles for survival analysis is based on a covariate-dependent version of the traditional Kaplan-Meier estimator, but is computationally impractical in many applied settings. We derive a continuous-analogue of the Kaplan-Meier estimator and describe how it can be used to do survival analysis with generic machine learning methods for conditional density estimation.

C0404: **Sensitivity to unobserved confounding in studies with factor-structured outcomes**
Presenter: **Alexander Franks**, UC Santa Barbara, United States
Co-authors: Alexander Alex

An approach is proposed for assessing sensitivity to unobserved confounding in studies with multiple outcomes. We demonstrate how prior knowledge unique to the multi-outcome setting can be leveraged to strengthen causal conclusions beyond what can be achieved from analyzing individual outcomes in isolation. We argue that it is often reasonable to make a shared confounding assumption, under which residual dependence amongst outcomes can be used to simplify and sharpen sensitivity analyses. We focus on a class of factor models for which we can bound the causal effects for all outcomes conditional on a single sensitivity parameter that represents the fraction of treatment variance explained by unobserved

confounders. We characterize how causal ignorance regions shrink under additional prior assumptions about the presence of null control outcomes, and provide new approaches for quantifying the robustness of causal effect estimates. Finally, we illustrate our sensitivity analysis workflow in practice, in an analysis of both simulated data and a case study with data from the National Health and Nutrition Examination Survey (NHANES).

CO023 Room BCB 310 COMPOSITIONAL, DISTRIBUTIONAL AND RELATIVE ABUNDANCE DATA Chair: Karel Hron

C0210: **Selection of relevant pairwise logratios for high-dimensional compositional data**
Presenter: **Paulina Jaskova**, Palacky University Olomouc, Czech Republic
Co-authors: Karel Hron, Javier Palarea-Albaladejo, Matthias Templ

In microbiome data analysis, one of the most important steps is to identify biomarkers. Microbiome data are characterized as high-dimensional compositional data, that is, relative data where the relevant information is contained in logratios between variables. Biomarkers usually can be represented by pairwise logratios which provide the key contained information. However, due to the high dimensionality of the data, it is a challenge, from a statistical perspective, to analyse all possible pairwise logratios since from each D-part composition, $D \cdot (D-1)/2$ pairwise logratios are derived. The main goal of this contribution is to present an algorithm to help us solve the problem of a high number of orthonormal logratio coordinate representations, which are needed for representation of individual logratios. The algorithm is based on the latin square theory for creating $D-1$ coordinate systems (balances), containing all logratios, which could be subsequently used in partial least squares regression to identify significant logratios. The properties of this new approach will be investigated using real high-dimensional compositions.

C0213: **Identification of important pairwise logratios in compositional data employing sparse principal component analysis**
Presenter: **Viktorie Nesrstova**, Palacky University, Olomouc, Czech Republic
Co-authors: Ines Wilms, Karel Hron, Peter Filzmoser

Compositional data are data that carry relative information as their elemental information is contained in the pairwise logratios of the parts that constitute the composition. While pairwise logratios are typically easy to interpret, the number of such possible pairs to consider quickly grows, thereby leading to a potentially exhaustive analysis even for medium-sized compositions. Sparse principal component analysis (PCA) therefore forms an appealing tool to identify important pairwise logratios, and in turn, the important parts in the composition. To this end, we apply the sparse PCA to the possibly high-dimensional matrix of all pairwise logratios. The L1 penalty in the optimization problem serves as tradeoff between explained variability and sparsity in the loadings of the pairwise logratios. The procedure is demonstrated on both simulated and empirical (geochemical) data sets. To aid practitioners in the discovery of important pairwise logratios, we introduce three practical visualization tools that (i) balance between the explained variability and sparsity of the model, (ii) show the stability of pairwise logratios, and (iii) highlight the importance of each particular part in the composition.

C0180: **Unsupervised changepoint detection for panel data**
Presenter: **Barbora Pestova**, Charles University, The Czech Academy of Sciences, Institute of Computer Science, Czech Republic
Co-authors: Michal Pesta, Matus Maciak

Detection procedures for a change in means of panel data are proposed. Unlike classical inference tools used for the changepoint analysis in the panel data framework, we allow for mutually dependent and generally non-stationary panels with an extremely short follow-up period. Two competitive self-normalized test statistics are employed, and their asymptotic properties are derived for a large number of available panels. Bootstrap extensions are introduced in order to handle such a universal setup. The presented unsupervised changepoint methods are able to detect a common break point even when the change occurs immediately after the first time point or just before the last observation period. The developed tests are proven to be consistent. Their empirical properties are investigated through a simulation study. The invented techniques are applied to option pricing.

C0293: **Approximation of bivariate densities with compositional splines**
Presenter: **Stanislav Skorna**, Palacky University, Czech Republic
Co-authors: Karel Hron, Jitka Machalova, Jana Burkotova

Multivariate densities occur often as a result of the aggregation of data in many applications. They are used to analyze the association structure and to further process them using methods of functional data analysis. For the purpose of further statistical analysis, proper spline (continuous) representation of the input discrete data is crucial. Bayes Hilbert spaces methodology enables to capture specific features of probability density functions and to construct so-called compositional splines which respect their decomposition into interactive and independent parts. Centered log-ratio, a key tool of this methodology, enables to represent the original densities (and compositional spline as as their estimates) in the standard L^2 space by *ZB*-spline representation. The resulting spline functions fulfill zero-integral condition, which must be taken into account already when building the basis of the *ZB*-spline representation. Basis can be built using standard *B*-spline basis with implemented zero-integral constraint or using the *ZB*-spline basis, which satisfies the zero-integral constraint automatically. We focus on the latter case, provide a detailed simulation study and apply the resulting spline representation for descriptive analysis of geochemical density data.

CO021 Room BCB 309 STATISTICS AND DATA ANALYTICS Chair: Stefan Van Aelst

C0317: **Non-parametric dimensionality detection for functional data**
Presenter: **Enea Bongiorno**, Universita del Piemonte Orientale, Italy
Co-authors: Kwo Lik Lax Chan, Aldo Goia

Several methodologies in functional statistics are based on multistep strategies, which, as a first step, involve a dimensionality reduction technique. Usually, a critical aspect of the latter technique is the choice of the number of components to use. We present a non-parametric strategy for this choice based on the notion of complexity for a process. This concept can be interpreted as a sort of degree of freedom of the process and is defined starting from an appropriate factorization for the Small Ball probability of the process itself. The methodology will be illustrated through simulations and applications.

C0354: **Multivariate finite-sample adjustments for equivalence testing**
Presenter: **Luca Insolia**, University of Geneva, Switzerland
Co-authors: Stephane Guerrier, Maria-Pia Victoria-Feser, Yanyuan Ma, Younes Boulaguiem, Dominique-Laurent Couturier

Average equivalence testing procedures aim at assessing whether two or more effects are comparable. They rely on the definition of a tolerance region inside which the compared effects (e.g., differences in means) could be considered negligible. This is in striking contrast to the traditional hypothesis testing framework, where the null and alternative hypotheses being tested are switched, as it reverses the burden of proof to demonstrate that the compared effects are indeed similar. The Two One-Sided Tests (TOST) procedure is widely used across different domains, but it is known for being too conservative. This leads to lower power to detect equivalence, especially in the presence of effects with higher variability. We propose a finite-sample adjustment of the TOST to guarantee that the resulting test is exactly of size alpha and, at the same time, is uniformly more powerful than existing methods based on the TOST. The proposed approach is defined at the population level, but it maintains good property when it is estimated from the data. Computationally lean algorithms, approximations, and extensions to multivariate equivalence testing problems are also discussed. Our results are supported by extensive Monte Carlo simulations and a real-world application related to pharmacokinetics bioequivalence.

C0294: **Fast linear model trees by PILOT**
Presenter: **Ruicong Yao**, KU Leuven, Belgium
Co-authors: Jakob Raymaekers, Peter Rousseeuw, Tim Verdonck
Linear model trees are regression trees that incorporate linear models in the leaf nodes. This preserves the intuitive interpretation of decision trees and, at the same time, enables them to better capture linear relationships, which is hard for standard decision trees. But most existing methods for fitting linear model trees are time-consuming and, therefore, not scalable to large data sets. In addition, they are more prone to overfitting and extrapolation issues than standard regression trees. We introduce PILOT, a new algorithm for linear model trees that is fast, regularized, stable and interpretable. PILOT trains in a greedy fashion like classic regression trees, but incorporates an L^2 boosting approach and a model selection rule for fitting linear models in the nodes. The abbreviation PILOT stands for PIecewise Linear Organic Tree, where 'organic' refers to the fact that no pruning is carried out. PILOT has the same low time and space complexity as CART without its pruning. An empirical study indicates that PILOT tends to outperform standard decision trees and other linear model trees on a variety of data sets. Moreover, we prove its consistency in an additive model setting under weak assumptions. When the data is generated by a linear model, the convergence rate is polynomial.

C0376: **Subset selection ensembles**
Presenter: **Stefan Van Aelst**, University of Leuven, Belgium
Co-authors: Anthony-Alexander Christidis, Ruben Zamar
Two key approaches for high-dimensional regression are sparse methods, such as best subset selection, and ensemble methods, such as random forests. Sparse methods have the advantage that they yield interpretable models. However, they are often outperformed in terms of prediction accuracy by blackbox multi-model ensemble methods. We propose an algorithm to optimize an ensemble of penalized regression models by extending recent developments in optimization for sparse methods to multi-model regression ensembles. The algorithm learns sparse and diverse models in the ensemble simultaneously from the data. Each of these models provides an explanation for the relationship between a subset of predictors and the response variable. To initialize our algorithm, forward stepwise regression is generalized to multi-model regression ensembles. The resulting ensembles achieve excellent prediction accuracy by exploiting the accuracy-diversity tradeoff of ensembles. The ensembles can outperform state-of-the-art competitors on both simulated and real data.

CC085 Room BCB 308 COMPUTATIONAL STATISTICS Chair: Mark De Rooij

C0307: **A polynomial-time algorithm for optimization-based depths**
Presenter: **Jeremy Guerin**, LTCI, Telecom Paris, Institut Polytechnique de Paris, France
Co-authors: Pavlo Mozharovskyi
Data depth is a statistical function which, by introducing centrality-based ordering, generalizes concepts of median and quantiles to higher dimensions. It has undergone substantial theoretical developments and is renowned for its attractive properties, such as affine invariance or robustness. In its variety of depth notions, data depth has become a universal methodology in statistics with numerous applications. However, using such functions in practice is limited by the computational complexity of algorithms which are exponential in data dimension. To cope with this computational intractability, we introduce a novel class of depth functions: depths that can be written as a polynomial (on a properly chosen domain) optimization problem. This class is sufficiently large to contain some of the most commonly used depth notions. In order to compute these depth functions, we suggest using a hierarchy of semidefinite programming relaxations. This method relies on the use of the sum-of-squares certificates of positivity. The goal is to obtain algorithms able to compute such functions in time that are polynomial in both size and dimension of the data set at hand. Finally, a simulation study explores in detail the properties of the proposed family of algorithms.

C0322: **Statistical estimation of heart movements by microwave Doppler radar sensor**
Presenter: **Takashi Ota**, Chuo University, Japan
Co-authors: Kosuke Okusa
The heart is a fundamental organ that sustains our life, and the development of a cardiac function monitoring method for daily use would offer significant benefits toward improving quality of life (QOL). Challenges in the use of home electrocardiographs include the wearing conditions of sensors, noise caused by body motion, and mental stress resulting from wearing sensors. We describe a new statistical method for estimating heart movements over time using a template method which is based on non-contact measurements with a microwave Doppler sensor and also employs mathematical modeling of heart movements, and simulation of Doppler reflected waveforms. Studies conducted using conventional Doppler sensors involve obtaining the subject's heartbeat reflected wave in advance, creating templates using statistical models and machine learning, and then measuring and analyzing the reflected wave again. This approach requires time-consuming diagnosis and is not suited for rapid analyses, such as streaming analysis. The results of monitoring the heart movements of multiple subjects using the new template method we proposed showed strong similarity with the observed reflected waves, suggesting the feasibility of our rapid and highly flexible heart movement estimation system.

C0324: **Efficient nonparametric two-sample testing with the maximum mean discrepancy**
Presenter: **Dean Bodenham**, Imperial College London, United Kingdom
Co-authors: Yoshinobu Kawahara
Many good nonparametric two-sample tests for univariate data exist, such as the Kolmogorov-Smirnov, Cramer-von Mises and Wilcoxon-Mann-Whitney tests. The maximum mean discrepancy (MMD) test is a nonparametric kernelised two-sample test that, when using a characteristic kernel, can detect any distributional change between two samples. It is defined for multivariate data, and when the total number of d-dimensional observations is n, direct computation of the test statistic is $O(dn^2)$. While approximations with lower computational complexity are known, more efficient methods for computing the exact test statistic are unknown. First, an exact method is described for computing the MMD test statistic for the univariate case in $O(n \log n)$ using the Laplacian kernel, and will compare its performance to classic univariate tests, highlighting cases where the MMD is more powerful. Moreover, this approach can be modified to create a $O(n \log n)$ algorithm for exactly computing the MMD statistic for bivariate data. For higher dimensions, the exact univariate method is extended to an approximate method, also with complexity log-linear in the number of observations. Experiments show that this approximate method can have good statistical performance when compared to the exact test, particularly in cases where $d > n$.

C0303: **Optimal control for parameter estimation in partially observed hypoelliptic stochastic differential equations**
Presenter: **Quentin Clairon**, ISPED - Universita de Bordeaux, France
Co-authors: Adeline Leclercq-Samson
The problem of parameter estimation in stochastic differential equations (SDEs) in a partially observed framework is considered. We aim to design a method working for both elliptic and hypoelliptic SDEs, the latter being characterized by degenerate diffusion coefficients. This feature often causes the failure of a contrast estimator based on the Euler Maruyama discretization scheme and dramatically impairs classic stochastic filtering methods used to reconstruct the unobserved states. All of these issues make the estimation problem in hypoelliptic SDEs difficult to solve. To overcome this, we construct a well-defined cost function no matter the elliptic nature of the SDEs. We also bypass the filtering step by considering a control theory perspective. The unobserved states are estimated by solving deterministic optimal control problems using numerical methods which

do not need strong assumptions on the diffusion coefficient conditioning. Numerical simulations made on different partially observed hypoelliptic SDEs reveal our method produces accurate estimates while dramatically reducing the computational price compared to other methods.

Friday 25.08.2023 09:00 - 10:00 Parallel Session M – COMPSTAT2023

CV035 Room BCB 206 COMPUTATIONAL AND FINANCIAL ECONOMETRICS Chair: Alessandra Luati

C0337: **Forecasting economic activity with a neural network in uncertain times: Application to German GDP**
Presenter: **Boris Kozyrev**, Halle Institute for Economic Research (IWH), Germany
Co-authors: Oliver Holtemoeller

The forecasting and nowcasting performance of a generalized regression neural network (GRNN) is analyzed. First, evidence from Monte Carlo simulations for the relative forecast performance of GRNN depending on the true but unknown data-generating process is provided. The analysis shows that GRNN outperforms autoregressive-moving average models in various practical scenarios. An additional check of fitting ARMA using simulated samples is provided. As a result, existing ARMA fitting approaches, even though in many cases yielding similar to GRNN predictions, often cannot properly identify a true DGP. Later, GRNN is applied to forecast quarterly German GDP growth with a distinction between "normal" times and situations with significantly different time-series behavior, such as during the COVID recession and recovery. The specific data transformation needs to be implemented, i.e., dividing aggregated level values of each indicator by the corresponding GDP value. Then, these ratios are used to perform one-step-ahead forecasting using GRNN. After that, using actual aggregated observations within a given quarter, a set of GDP nowcasts is obtained. This algorithm has a high forecasting power, outperforming traditional nowcasting models (AR(1), DFM, model averaging), especially during the COVID-19 crisis.

C0160: **Informative priors to estimate the value-at-risk**
Presenter: **Mario M Pizarro**, Universidad de Extremadura, Spain
Co-authors: Eva Lopez Sanjuan, M Isabel Parra Arevalo

In Risk Theory, the use of the tools provided by Extreme Value Theory is essential to estimate risk measures, but usually, only the observations that exceed a certain fixed value are taken into account for the estimation. Value at Risk (VaR) and Conditional Value at Risk (CVaR) are the most employed risk measures. A new Bayesian method, based on Metropolis-Hastings (MH) algorithm, is proposed in order to estimate VaR. The method employs informative prior distributions for the parameters of the Generalized Pareto distribution (GPD), using all the datasets and consequently, seizing all the information available. In order to compare the quality of the estimates provided, a broad simulation study is carried out for different distributions of the data. The results show that the new strategy provides better estimates for VaR than standard MH with non-informative priors.

C0411: **Factor-augmented sparse MIDAS regression for nowcasting**
Presenter: **Jonas Striaukas**, Copenhagen Business School, Denmark

GDP nowcasting commonly employs either sparse regression or a dense approach based on factor models, which differ in the way they extract information from high-dimensional datasets. This paper aims to investigate whether sparse regression of the outcome on both the covariables and the factors can improve nowcasts. We propose an estimator for a factor-augmented sparse regression model. The rates of convergence of the estimator are derived in a time series context, accounting for tau-mixing processes and fat-tailed distributions.The application of this new technique to nowcast US GDP growth reveals several key findings. Firstly, our sparse plus dense technique significantly improves the quality of nowcasts compared to both sparse and dense benchmarks over a period period from 2008 Q1 to 2022 Q2. This improvement is particularly pronounced during the COVID pandemic, indicating the model's ability to capture the specific dynamics introduced by the pandemic. Interestingly, our novel factor-augmented sparse method does not perform significantly better than sparse regression prior to the onset of the pandemic, suggesting that using only a few predictors is sufficient for nowcasting in more stable economic times.

CO014 Room BCB 309 RECENT CLUSTERING METHODS FOR COMPLEX DATA I Chair: Mika Sato-Ilic

C0219: **Mixture of linear mixed models for clustering weighted random graphs**
Presenter: **Shu-Kay Angus Ng**, Griffith University, Australia
Co-authors: Richard Tawiah, Hien Nguyen, Florence Forbes

Typical clustering methods assume observed data are independent. However, this assumption is often not valid with modern data (e.g., random graphs consisting of a set of nodes and a relational tie measured on each pair of nodes). Clustering methods that are not adequate to capture the complex dependence structure among highly correlated data often lead to biased estimates, misleading conclusions, lack of model fit, unstable inference, and inaccurate presentation of heterogeneity or data variability. The aim is to develop a new statistical approach using mixtures of linear mixed models (LMMs) for clustering weighted random graphs (where observed edge responses are real numbers). Random effects models have been a widely successful tool in capturing complex correlations among observations. Building on the developments in mixture models and LMMs, the proposed approach incorporates two sets of random effects in the linear predictor for the mean response to capture within-node and transitivity dependences and model node-level and paired node-level variability. Maximum likelihood estimation of the unknown parameters can be performed conditionally on the node-specific random effects within an incomplete-data framework of the expectation-maximisation (EM) algorithm. The proposed method is applied in comorbidity research using a real-world data set from the Australian National Health Survey (NHS) to identify clusters of comorbid medical conditions.

C0235: **Asymmetric cluster difference scaling based on hill-climbing model**
Presenter: **Kensuke Tanioka**, Doshisha University, Japan
Co-authors: Hiroshi Yadohisa

Asymmetric (dis)similarity data is dissimilarity data such that dissimilarity from subject i to subject j does not necessarily match the dissimilarity from subject j to i. In fact, asymmetric (dis)similarity data is observed in various situations such as brand switching, or network analysis for SNS. Given asymmetric (dis)similarity data, Asymmetric Multidimensional Scaling (AMDS) is a very useful tool to interpret the asymmetric relations between subjects visually. AMDS provides us with two things on the lower dimension; One is the coordinates of each subject, and the other is parameters for describing asymmetric relations. There are various AMDS have been proposed, and these methods depend on the model of parameters for the asymmetric relation. However, these days, information technology is improved, and we have to deal with large and complex data. If AMDS is applied to such large asymmetric (dis)similarity data, it becomes difficult to interpret the asymmetric relation because the number of subjects is large. To overcome the problem, we propose new AMDS for the cluster centroids, not subjects. The proposed AMDS visualizes asymmetric relations between clusters. In the proposed method, the hill-climbing model is adopted from the candidates of AMDS models.

C0236: **A fuzzy cluster-scaled principal component analysis for mixed high-dimension and low-sample size data**
Presenter: **Mika Sato-Ilic**, University of Tsukuba, Japan

A fuzzy cluster-scaled principal component analysis (fuzzy cluster-scaled PCA) for mixed high-dimension and low-sample size data (mixed HDLSS) is proposed. The mixed HDLSS data comprises numerical and categorical data regarding quantitative and qualitative variables, and the number of variables is much larger than the number of objects. Conventionally, fuzzy cluster-scaled PCA has been proposed to analyze HDLSS data because ordinary PCA cannot be applied to HDLSS data since, theoretically, we cannot obtain the correct solution because of ordinary PCA. The essence of the fuzzy cluster-scaled PCA is the utilization of the result of fuzzy clustering. Fuzzy cluster-scaled PCA is based on the fuzzy cluster-scaled correlation, which is decomposed into two parts. First is the correlation of classification structures obtained because of fuzzy clustering, and this can be obtained using the dissimilarity of categorical data regarding qualitative variables. Second, is the ordinary correlation between

variables, so we can use numerical data regarding quantitative variables. For constructing the fuzzy cluster-scaled correlation, the two parts used different kinds of data, which are numerical and categorical data, are reasonably combined, and we can obtain the result of PCA for the mixed HDLSS data. Several numerical examples show a better performance of the proposed method.

CC045 Room BCB 307 APPLIED STATISTICS AND DATA ANALYSIS Chair: Luca Insolia

C0150: **Federated learning via distributed sequential method**
Presenter: **Yuan-chin Ivan Chang**, Academia Sinica, Taiwan

The analysis of data stored in multiple sites has become more popular, raising new concerns about the security of data storage and communication. Federated learning, which does not require centralizing data, is a common approach to preventing heavy data transportation, securing valued data, and protecting personal information protection. Therefore, determining how to aggregate the information obtained from the analysis of data in separate local sites has become an important statistical issue. The commonly used averaging methods may not be suitable due to data nonhomogeneity and incomparable results among individual sites, and applying them may result in the loss of information obtained from the individual analyses. Using a sequential method in federated learning with distributed computing can facilitate integration and accelerate the analysis process. We develop a data-driven method for efficiently and effectively aggregating valued information by analyzing local data without encountering potential issues such as information security and heavy transportation due to data communication. In addition, the proposed method can preserve the properties of classical sequential adaptive design, such as data-driven sample size and estimation precision when applied to generalized linear models. We use numerical studies of simulated data and an application to COVID-19 data collected from 32 hospitals in Mexico, to illustrate the proposed method.

C0299: **Composite lognormal distributions of cosmic voids in simulation and mock data**
Presenter: **Nour Hamed**, American University of Sharjah, United Arab Emirates
Co-authors: Stephen Chan

The aim is to illustrate the power of the composite lognormal distributions and to show that it provides consistently better fits than the commonly used three-parameter lognormal distribution for modeling void size distributions of the Cosmic Void Catalog (CVC) based on three different simulation and mock catalogs; dark matter, haloes and galaxies. The results will be truly beneficial for applications of composite lognormal distributions in other areas of astronomy and astrophysics. Moreover, they can potentially assist and play a crucial role in understanding the dynamical processes affecting the structure and formation of the Universe.

CC118 Room BCB 308 QUALITY CONTROL Chair: Steven Gilmour

C0177: **Tolerance interval and control chart for mixture distribution**
Presenter: **Hsiuying Wang**, National Yang Ming Chiao Tung University, Taiwan

Tolerance intervals have wide applications in various industries, including manufacturing engineering, clinical research, and pharmaceuticals. In some manufacturing processes, defects can arise from multiple factors. In such cases, using a mixture distribution of suitable probabilistic models can be more appropriate than a simple model. The tolerance intervals for the normal mixture distribution have been studied. Tolerance intervals can also be used to establish control charts to monitor quality characteristics. An approach to constructing modified two-sided tolerance intervals for the normal mixture distribution has been proposed in the literature. The procedure for using the modified tolerance intervals is presented. It summarizes a rule for constructing the modified two-sided tolerance intervals of the normal mixture distribution, which can be extended to construct two-sided tolerance intervals for general mixture distributions. Additionally, the feasibility of using tolerance intervals to develop control charts for mixture distributions is discussed.

C0176: **A new phase II change point detection control chart for monitoring and diagnostics of linear profiles**
Presenter: **Longcheen Huwang**, National Tsing Hua University, Taiwan

A new Phase II control chart, which is based on the change point model and combined with the exponentially weighted moving average (EWMA) mechanism, is proposed to monitor general linear profiles. The new control chart can be used to monitor general linear profiles when the true in-control parameters are unknown and only a few historical data are available. In addition, when the chart triggers an out-of-control signal, it cannot only estimate the location of the change point, but it can also identify which of the parameters have changed and the change directions. Using Monte-Carlo simulations, the proposed chart is shown to be effective and has good diagnostic performance. Furthermore, the simulation results show that the proposed chart performs better than the existing charts in most out-of-control scenarios considered. An example is used to illustrate how the proposed chart can be implemented in practical applications.

C0352: **Rectifying LTPD sampling plans for combined inspection by variables and attributes**
Presenter: **Jindrich Klufa**, Prague University of Economics and Business, Czech Republic

The acceptance sampling is studied. We shall consider the lot tolerance percent defective (denoted LTPD) sampling inspection plans when the remainder of a rejected lot is inspected, i.e. the rectifying LTPD plans. These plans were introduced by Dodge and Romig for inspection by attributes (each inspected item is classified as either good or defective). Similar rectifying plans for inspection by variables were introduced by the author of this contribution. We shall consider combined inspection in which all items from the sample are inspected by variables, but the remainder of a rejected lot is inspected only by attributes. We shall show that the combined inspection is optimal in many practical situations. Using the LTPD plans for combined inspection by variables and attributes, we can often achieve significant savings in the inspection cost under the same protection of producer and consumer.

CC070 Room BCB 310 TEXT MINING Chair: Maria Brigida Ferraro

C0281: **Visualizing topic uncertainty in topic modelling**
Presenter: **Peter Winker**, University of Giessen, Germany

Word clouds became a standard tool for presenting results of natural language processing methods such as topic modelling. They exhibit the most important words, where word size is often chosen proportional to the relevance of words within a topic. In the latent Dirichlet allocation (LDA) model, word clouds are graphical presentations of a vector of weights for words within a topic. These vectors are the result of a statistical procedure based on a specific corpus. Therefore, they are subject to uncertainty coming from different sources such as sample selection, random components in the optimization algorithm, or parameter settings. A novel approach for presenting word clouds, including information on such types of uncertainty, is introduced and illustrated with an application of the LDA model to conference abstracts.

C0297: **Choosing the number of topics in LDA models: A Monte Carlo comparison of selection criteria**
Presenter: **Anna Staszewska-Bystrova**, University of Lodz, Poland
Co-authors: Victor Bystrov, Viktoriia Naboka, Peter Winker

Selecting the number of topics in LDA models is considered to be a difficult task, for which alternative approaches have been proposed. The performance of the recently developed singular Bayesian information criterion (sBIC) is evaluated and compared to the performance of alternative model selection criteria. The sBIC is a generalization of the standard BIC that can be implemented in singular statistical models. The comparison is based on Monte Carlo simulations and carried out for several alternative settings, varying with respect to the number of topics, the number of

documents and the size of documents in the corpora. The performance is measured using different criteria which take into account the correct number of topics, but also whether the relevant topics from the DGPs are identified. Practical recommendations for LDA model selection in applications are derived.

C0331: **Textual content and academic journals selectiveness: A case of economic journals**
Presenter: **Pawel Baranowski**, Institute of Economic and Financial Research, Lodz, Poland, Poland
Co-authors: Szymon Wojcik

Currently observed vast influx of papers obstructs the editorial procedures in scientific journals. This phenomenon applies explicitly to top-quality academic journals with high scientific impact. Moreover, it stimulates the emergence of low- (or non-) selective journals, attracting authors with short editorial procedures in exchange for high fees. We argue that introducing natural language processing (NLP) can help distinguish the papers worth reading by the editor from those whose scientific quality does not meet the standards of the journal. To test this hypothesis, we apply state-of-art large language models, i.e. bidirectional encoder representations from transformers (BERT). Our sample consists of approximately 400 academic papers representing economics, finance or business. The papers were collected from journals of three levels of selectiveness, namely: highly selective (top-tier journals), moderately selective (journals listed on DOAJ list), and non-selective ("predatory" journals). More specifically, we used a pre-trained Sci-BERT model on anonymized and pre-processed texts of academic papers. The results show that the pure textual content may give more than 80% out-of-sample accuracy in classifying texts into the three levels of selectiveness. The outcomes of the study prove the usefulness of NLP in distinguishing the scientific quality of the paper and support Beall's classification of "predatory" journals.

CC094 Room BCB 311 LONGITUDINAL AND FUNCTIONAL DATA ANALYSIS Chair: Sonja Greven

C0220: **The mean group estimators for multi-level autoregressive models with intensive longitudinal data**
Presenter: **Kazuhiko Hayakawa**, Hiroshima University, Japan
Co-authors: Boyan Yin

The mean group (MG) estimators are proposed to estimate multilevel (vector) autoregressive models with intensive longitudinal data. The MG estimator is originally proposed in econometrics but is new to the behavioral science literature. Since the naive MG estimator suffers from the small sample bias problem, jackknife and analytical bias corrections are proposed. It is argued that the MG estimator has several advantages over existing methods, such as restricted maximum likelihood or Bayesian methods in terms of model specification and implementation. Monte Carlo simulation is performed to investigate the performance of the MG estimators and compare them with the existing methods. The simulation results indicate that the bias-corrected MG estimators have superior or comparable performance compared to the existing methods.

C0355: **General estimation framework for multi-state Markov processes with flexible specification of the transition intensities**
Presenter: **Alessia Eletti**, University College London, United Kingdom
Co-authors: Giampiero Marra, Rosalba Radice

When interest lies in the progression of a disease rather than in a single outcome, non-homogeneous multi-state Markov models constitute a natural and powerful modelling approach. Constant monitoring of a phenomenon of interest is often unfeasible, hence leading to an intermittent observation scheme. This setting is challenging and existing models and their implementations do not yet allow for flexible enough specifications that can fully exploit the information contained in the data. To widen the scope of multi-state Markov models significantly, we propose a closed-form expression for the local curvature information of a key quantity, the transition probability matrix. Such development allows one to model any type of multi-state Markov process, where the transition intensities are flexibly specified as functions of additive predictors. Parameter estimation is carried out through a carefully structured, stable penalised likelihood approach. The methodology is exemplified via two case studies that aim at modelling the onset of cardiac allograft vasculopathy and cognitive decline. To support applicability and reproducibility, all developed tools are implemented in the R package flexmsm.

C0405: **Reconstructing partially observed functional data via factor models of increasing rank**
Presenter: **Maximilian Ofner**, Graz University of Technology, Austria
Co-authors: Siegfried Hoermann

In functional data analysis, applied researchers often face the problem of missing fragments. To recover the missing information from the observed parts, linear reconstruction operators have been introduced in the literature. In this setting, we present a new approach for estimating linear reconstructions using approximate factor models of increasing rank. The proposed methodology aims at discretely sampled functional data with additive noise and avoids restrictive smoothness conditions. Under a triple asymptotic, we establish uniform convergence rates of our estimator. Furthermore, we discuss a simple and effective method for constructing simultaneous prediction bands. Finite sample properties of the proposed procedures are then examined in a simulation study. The methodology is finally illustrated by a set of incompletely observed temperature data.

Friday 25.08.2023 10:30 - 12:00 Parallel Session N – COMPSTAT2023

CV031 Room BCB 308 TIME SERIES AND DEPENDENCE MODELS Chair: Alessandra Luati

C0344: **Mean stationarity test in time series: A signal variance-based approach**
Presenter: **Kin Wai Chan**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Hon Kiu To

The inference of mean structure is an important problem in time series analysis. Various tests have been developed to test for different mean structures, for example, the presence of structural breaks, and parametric mean structures. However, many of them are designed for handling specific mean structures, and may lose power upon violation of such structural assumptions. We propose a new mean stationarity test built around the signal variance. The proposed test is based on a super-efficient estimator which could achieve a convergence rate faster than $\sqrt{n}$. It can detect the non-constancy of the mean function under serial dependence. It is shown to have promising power, especially in detecting hardly noticeable oscillating structures. The proposal is further generalized to test for smooth trend structures and relative signal variability. A real-data application on global land surface temperature data is presented. This research was partially supported by General Research Fund 14304420, 14306421, and 14307922 provided by the Research Grants Council of HKSAR.

C0342: **A vector error correction model to address sensor-based time series**
Presenter: **Maria del Carmen Robustillo Carmona**, Universidad de Extremadura, Spain
Co-authors: Lizbeth Naranjo Albarran, M Isabel Parra Arevalo, Carlos Javier Perez Sanchez

Vector Error Correction (VEC) models are useful for analyzing complex and dynamic long-term relationships between variables under a cointegration approach. These models have been widely used in some areas, such as econometrics, but in a very limited way to time series of sensor data, where they could be very useful. Two particularly interesting cases are i) sensor data from variables related to the status of the beehives in the context of precision beekeeping; ii) sensor data from acoustic features extracted from voice recordings in the context of computer-aided diagnosis systems for detecting Parkinsons' disease. Several experiments have been conducted to assess the performance of this model, and comparisons with linear multivariate state-space models have been made. A precision beekeeping dataset obtained from the we4bee database is shown as an example. Four inner temperatures, humidity, and weight, regularly collected from four beehives, led to four multivariate time series. Mean absolute errors of these variable predictions were estimated for 1, 3, and 7 days in a rolling cross-validation framework. Overall, the VEC model provided better results between 83.3% and 92.6% of the predictions made, depending on the beehive. The achieved improvement percentages suggest that considering cointegration in sensor data contexts may provide more competitive models.

C0371: **Exploring the impact of non-linear dependencies in stock market returns regime transitions**
Presenter: **Marina Dolfin**, King's College London, United Kingdom
Co-authors: Jose De Leon Miranda

Stock markets exemplify complex systems characterized by non-linear dependencies among actors at different levels of observation: at a micro level, concerning individual stocks, and at a macro level, when multiple markets interact. Importantly, regime transitions in these complex systems may endogenously arise due to the non-linear interactions, often shaped by the inherent heterogeneity among the market actors, resulting in asymmetrical correlations. Based on these considerations, we explore the impact of non-linear dependencies in the dynamics of market returns by analysing their time-dependent cross-correlations. We employ and compare several approaches, including Detrended Cross-Correlation Analysis (DCCA), spatial correlations, autocorrelations (to detect bifurcations) and minimum spanning trees. The principal objective of our research is to discern patterns that emerge during regime phase transitions and to identify potential early warning signals preceding regime changes, including market crashes. Additionally, we propose investigating the network system's controllability as a future research perspective. The data under analysis comprises returns from stock indices across different geographical regions.

C0375: **Consistent estimation of multiple breakpoints in dependence measures**
Presenter: **Marvin Borsch**, Institut for Okonometrie und Statistik, Germany
Co-authors: Alexander Mayer, Dominik Wied

Different methods are proposed to consistently detect multiple breaks in copula-based dependence measures. We allow for breaks in multiple and grouped dependence measures. Starting with the classical binary segmentation, also the more recent wild binary segmentation (WBS) is considered. For binary segmentation, consistency of the estimators for the location of the breakpoints as well as the number of breaks is proved, taking filtering effects from AR-GARCH models explicitly into account. Monte Carlo simulations based on a factor copula as well as on a Clayton copula model illustrate the strengths and limitations of the procedures. A real data application on recent Euro Stoxx 50 data considering the COVID-19 pandemic reveals some interpretable breaks in the dependence structure.

CI007 Room BCB 310 RECENT ADVANCES IN DIMENSION REDUCTION METHODS Chair: Sara Taskinen

C0192: **An adaptive approach for sparse quantile regression**
Presenter: **Andreas Artemiou**, University of Limassol, Cyprus
Co-authors: Christou Antonis

A new approach is presented for the penalization of the quantile regression. We propose an iterative procedure which is based on an approximation of the L_0 penalty, and the estimation involves the solution of a quadratic programming optimization problem. We compare our results with the LASSO quantile regression implemented in the quantreg package in R and demonstrate the advantages of our methodology.

C0196: **Tandem clustering with ICS**
Presenter: **Klaus Nordhausen**, University of Jyvaskyla, Finland
Co-authors: Andreas Alfons, Aurore Archimbaud, Anne Ruiz-Gazen

Tandem clustering is a well-known technique for dealing with high-dimensional or noisy data to better identify clusters. It is a sequential approach based on first reducing the dimension of the data and then performing the clustering. The most common method, based on principal component analysis (PCA), has been criticized for only focusing on maximizing inertia and not necessarily preserving the structure of interest for clustering. Therefore, we suggest a new tandem clustering approach based on invariant coordinate selection (ICS). This multivariate method is designed to identify the structure of the data by jointly diagonalizing two scatter matrices. More specifically, some theoretical results proved that under some elliptical mixture models, the first and/or last components carry the information regarding the clustering structure. The issues of choosing the pair of scatter matrices and the components to keep are the two challenges that must be addressed. For clustering purposes, we suggest that the best scatter pairs consist of one matrix which captures the within-cluster structure and another which captures the global structure. To this end, the local shape or pairwise scatters prove to be good choices for estimating the within-structure. The performance of ICS as a dimension reduction method is evaluated to determine its ability to preserve the cluster structure of the data.

C0229: **Spatio-temporal coregionalization modeling by using simultaneous diagonalization**
Presenter: **Sandra De Iaco**, University of Salento, Italy

The spatio-temporal linear coregionalization model (ST-LCM) represents one of the most common models applied to describe the correlation of

multiple variables which evolve in space-time. Thanks to its computational flexibility, it has been recalled in several studies and some advances support a simplified modeling stage through simultaneous diagonalization of the covariance matrices estimated for different lags as well as the choice of appropriate basic covariance models at the different spatio-temporal variability scales. Without these developments, the detection of the uncorrelated components through the identification of the nested structures from the empirical direct and cross-covariance functions would be a hard step in a spatio-temporal context, since 3D plots must be analyzed. Moreover, the selection of the same class of covariance models for all basic hidden components, usually proposed in the past, can be overcome by enabling each basic component to be modelled based on its own features and then by fitting a proper class of covariance models. The ST-LCM fitting process and some computational tools which improve the definition of the uncorrelated components and the main characteristics of the empirical covariance surfaces of the uncorrelated components (in terms of symmetry, separability/non-separability, type of non-separability) are presented together with an application.

CO010 Room Virtual room R01 RECENT ADVANCES IN BAYESIAN ECONOMETRICS Chair: Toshiaki Watanabe

C0274: **Analyzing intraday variation in price impact: A Bayesian SVAR approach with stochastic volatility estimation**
Presenter: **Makoto Takahashi**, Hosei University, Japan

The aim is to analyze the intraday variation in the short- and long-term price impact of market orders, limit orders, and cancellations using a structural vector autoregression (SVAR) model. While Bayesian estimation using sign restrictions has been effective in parameter estimation in SVAR models, there are issues with parameter uniqueness. To address this, alternative methods like maximum likelihood estimation and generalized method of moments have been proposed. This study applies a new Bayesian estimation method that considers the stochastic volatility of errors to estimate the model parameters. This method allows the unique identification of the parameters without being affected by the order of the variables by imposing sign conditions on the variables in addition to the heteroskedasticity of the variables. The advantage of this method is that the sign conditions can be easily verified from the posterior distribution of the estimated parameters. This estimation method has not been applied to high-frequency order book data, but the use of a large number of observations allows the model to be estimated every few minutes to tens of minutes and examine the intraday variation. The model simultaneously analyzes the variation in price impact and volatility by modeling and estimating the stochastic variation of both price changes and orders.

C0277: **Time-varying parameter local projections with stochastic volatility**
Presenter: **Jouchi Nakajima**, Hitotsubashi University, Japan

The local projection method has been widely used as a promising framework for computing impulse responses. In the previous literature, a time-varying version of the local projection has been proposed, but it does not address the time-varying variance of the error distribution. Ignoring a possible time variation in the error variance could cause a severe bias in the time-varying impulse responses. To overcome it, the time-varying parameter local projections with stochastic volatility are proposed. A Bayesian method using an efficient Markov chain Monte Carlo is developed to analyze the proposed model. The application to monetary policy effectiveness is provided using the U.S. macroeconomic variables.

C0279: **Posterior inferences on incomplete structural models: The minimal econometric interpretation**
Presenter: **Takashi Kano**, Hitotsubashi University, Japan

The minimal econometric interpretation (MEI) of DSGE models provides a formal model evaluation and comparison of misspecified nonlinear dynamic stochastic general equilibrium (DSGE) models based on atheoretical reference models. The MEI approach recognizes DSGE models as incomplete econometric tools that provide only prior distributions of targeted population moments but have no implications for actual data and sample moments. Based on the MEI approach, develops a Bayesian posterior inference method is developed. Prior distributions of targeted population moments simulated by the DSGE model restrict the hyperparameters of Dirichlet distributions. These are natural conjugate priors for multinomial distributions followed by corresponding posterior distributions estimated by the reference model. The Polya marginal likelihood of the resulting restricted Dirichlet-multinomial model has a tractive approximated log-linear representation of the Jensen-Shannon divergence, which the proposed distribution-matching posterior inference uses as the limited information likelihood function. Monte Carlo experiments indicate that the MEI posterior sampler correctly infers calibrated structural parameters of an equilibrium asset pricing model and detects the true model with posterior odds ratios.

C0278: **Time-varying macroeconomic announcement risk**
Presenter: **Jonathan Stroud**, Georgetown University, United States
Co-authors: Michael Johannes, Norman J Seeger

An issue overlooked in the finance and economics literature is examined: time variation in announcement volatility or event risk. We combine long spans of high-frequency data with a flexible parametric model of returns, which allows us to identify announcement returns, capture intraday volatility dynamics, and identify conditional announcement volatility. Long time spans are needed due to the infrequency of most announcements. We focus on crude oil due to its economic importance, high volatility and complex announcement structure. Results indicate strong evidence for time-varying announcement volatility, as announcement event risk varies by as much as an order of magnitude over time.

CO022 Room BCB 307 STATISTICS APPLIED TO INDUSTRY Chair: Francisco Louzada

C0338: **Reliability in Brazil: Roads for approaching industry**
Presenter: **Francisco Louzada**, University of Sao Paulo, Brazil

Our dependence on mechanical and electronic devices is increasing. However, no matter how efficient they are, they can fail. For instance, we can mention technologies embedded in intelligent sensors, artificial intelligence devices, agricultural, financial, and medical robots. In this context, statistical reliability analysis has been extensively used and inserted into innovation processes. Some reliability innovation projects are presented, showing how we are creating a connection road between academia and the industrial, medical and financial sectors. The focus is on reliability modeling for oil well construction equipment, bucket tracking equipment, agricultural machinery, and communication modeling for mobile phones.

C0343: **Statistical modeling and reliability analysis of repairable systems with dependent failure times under imperfect repair**
Presenter: **Paulo Ferreira**, Federal University of Bahia, Brazil
Co-authors: Eder Brito, Vera Tomazella, Francisco Louzada, Oilson Gonzatto-Junior

Imperfect repairs (IRs) are widely applicable in reliability engineering since most equipment is not completely replaced after failure. In this sense, it is necessary to develop methodologies that can describe failure processes and predict the reliability of systems under this type of repair. One of the challenges in this context is to establish reliability models for multiple repairable systems considering the dependency and/or unobserved heterogeneity between systems and the times of their respective failures after performing IRs. Thus, frailty models are proposed to identify these failure processes' statistical dependence and unobserved heterogeneity. In this context, we consider the arithmetic reduction of age (ARA) and arithmetic reduction of intensity (ARI) classes of IR models, with constant repair efficiency, a power-law process distribution to model failure times, and a shared gamma distributed frailty by all systems. Classical inferential methods are used to estimate the parameters and reliability predictors of systems under IRs. An extensive simulation study is carried out under different scenarios to investigate the suitability of the models and the asymptotic consistency and efficiency properties of the maximum likelihood estimators. We illustrate the practical relevance of the proposed models on two real data sets.

C0276: **Stats in Industry 5.0: Some cases of contemporaneous experimental designs adopting dynamic and hierarchical structures**
Presenter: **Diego Nascimento**, Universidad de Atacama, Chile

The advancement of technology has increased competitiveness, especially in the manufacturing industry. Given the competitiveness of the business world, decision-making based on data can support companies in fast and accurate strategic planning. Using statistical solutions, preliminary results about a novel water extraction in the worlds driest non-polar desert are presented. This region is a mining area in which the field demands a high volume of water. Therefore, a hierarchical dynamic spatial-temporal model was developed to estimate the relative humidity flux across the region. Then, Statistical Process Control (SPC) tools and Experimental Designs were adopted to draw fresh water in the form of thick cloud banks (known as Camanchaca). Results show that strategies adopting statistical solutions for capturing this moisture are arguably revolutionizing modern desert water collection.

C0407: **Statistical inference for generalized power-law process in repairable systems**
Presenter: **Pedro Ramos**, Pontificia Universidad Catolica de Chile, Brazil

Repairable systems are often used to model the reliability of restored components after a failure is observed. Among various reliability growth models, the power law process (PLP) or Weibull process has been widely used in industrial problems and applications. We propose a new class of model called generalized PLP (GPLP), based on change points, which can be treated as known or unknown parameters or interpreted as failure times, in which we consider the impact of all or some fixes about fault intensity function. In this context, GPLP, unlike the usual PLP, is not restricted to the assumption of minimal repair (RM), it is possible to consider other situations, such as perfect, efficient, and harmful repair. Some special cases of GPLP are presented, such as the main models for the analysis of repairable systems under the assumption of imperfect repair. The estimators of the proposed model were obtained using the maximum likelihood method. We evaluated the performance of the parameter estimators through Monte Carlo (MC) simulations. The proposed approach is fully illustrated two real failure time datasets.

CO018 Room BCB 311 ML AND FINTECH Chair: Maria Grith

C0205: **Posterior contraction for deep Gaussian process priors**
Presenter: **Gianluca Finocchio**, University of Vienna, Austria
Co-authors: Johannes Schmidt-Hieber

Posterior contraction rates are studied for a class of deep Gaussian process priors in the nonparametric regression setting under a general composition assumption on the regression function. It is shown that the contraction rates can achieve the minimax convergence rate (up to $\log n$ factors), while being adaptive to the underlying structure and smoothness of the target function. The proposed framework extends the Bayesian nonparametric theory for Gaussian process priors.

C0238: **On pricing kernels for digital assets**
Presenter: **Ratmir Miftachov**, Humboldt University of Berlin, Germany
Co-authors: Maria Grith, Zijin Wang

Time-varying preferences and the risk aversion of digital market investors are investigated by estimating pricing kernels (EPK) on Bitcoin (BTC) options data. We compare the classical method based on risk-neutral and physical density to the so-called conditional density integration (CDI) method that incorporates forward-looking information and B-spline functions. The CDI estimator takes into account the information available to investors but misses a proper construction of confidence bands that allow comparing EPKs estimated by both methods. Further, we use a functional principal component approach on the latter B-spline coefficients to cluster the estimated pricing kernels. The clusters are interpreted as different regimes and set in relation to other financial market measures. Our empirical results show that short- and long-term investors differ significantly regarding their EPKs.

C0260: **Modeling nonlinear dynamics of functional time series for large-scale data**
Presenter: **Hannah Lan Huong Lai**, National University of Singapore, Singapore
Co-authors: Maria Grith, Ying Chen

In numerous empirical applications, financial and economic data can be naturally represented as curves or surfaces exhibiting nonlinear dynamics. To address this challenge, we propose a Nonlinear Functional Autoregressive model (NFAR) that leverages neural networks to capture the nonlinear serial dependence of functional time series data. The estimation of NFAR models involves a two-stage procedure. In the first stage, we extract low-dimensional components from the covariance operator of the response and explanatory variables. In the second stage, we use neural networks to approximate the complex and nonlinear patterns in the data. We further explore the asymptotic properties of the NFAR models. To demonstrate the effectiveness of our proposed approach, we apply the NFAR model to the daily implied volatility surfaces of the S&P 500 index options from 2009 to 2021. Our results showcase superior prediction accuracy and substantial economic gains, which we illustrate via several trading strategies. These findings suggest the potential of our proposed method in capturing complex dynamics of financial and economic data, thereby providing valuable insights for investors.

C0264: **Spectral factors for functional data**
Presenter: **Maria Grith**, Erasmus University Rotterdam, Netherlands

A stylized fact has emerged that volatility, skewness, kurtosis, and term structure factors effectively forecast future implied volatility surfaces. These surfaces reflect investors' anticipated market conditions at various points in the future. A novel approach is proposed for disentangling risk factors that capture fluctuations across cycles of different lengths. Specifically, we adopt a double orthogonal decomposition of the implied volatility surfaces in the time and space domains. Our method allows us to estimate frequency-specific risk factors that are spectral counterparts of those commonly identified in the existing literature. These spectral factors offer valuable insights into the behavior of various investors operating under distinct market conditions and may reduce dimensions in the factor space in an economically-meaningful way. In addition, our approach demonstrates the potential to improve the accuracy of forecasting implied volatility curves relative to the traditional methods.

CO020 Room BCB 309 RECENT CLUSTERING METHODS FOR COMPLEX DATA II Chair: Mika Sato-Ilic

C0249: **Clustering for category variables in linear regression via generalized fused Lasso**
Presenter: **Mineaki Ohishi**, Tohoku University, Japan
Co-authors: Hirokazu Yanagihara

In linear regression, we often use category variables as explanatory variables. A category variable has two types: one is a qualitative variable and the other one is obtained by splitting a quantitative variable. Regarding the former, the obtained finest categories are usually used for modelling. The latter is a popular way of modeling some sort of value, such as real estate. Moreover, the use of the latter has the merit that a non-linear structure can be naturally incorporated into a linear model. When using the latter, a quantitative variable is usually split into divisions based on some experience. However, too fine categories may cause overfitting and complicate a model's interpretation. On the other hand, unsuitably clustered categories may cause declining model fitting. Hence, it is important to consider optimizing the cluster of categories. To address this, we develop an estimation method involving clustering of categories via generalized fused Lasso. Using categories as fine as possible, by estimating parameters for categories with similar effects to be exactly equal, we can expect to obtain the optimal cluster of categories.

C0250: **Comparison of prediction methods for spatial data using real estate data**

Presenter: **Koki Kirishima**, Hiroshima University, Japan
Co-authors: Mineaki Ohishi, Hirokazu Yanagihara

Data with location information such as latitude and longitude are called spatial data. We consider a prediction problem for real estate value which is a typical example of spatial data. As a method of spatial data analysis, Geographically Weighted Regression and Spatially Clustered Regression have been proposed. In addition, an estimation method using the adjacency relations in the space by Generalized Group Fused LASSO has also been proposed. The second and third methods can also be used to cluster spatial data. By applying these methods to real estate data, we compare the accuracy of predictions based on each method. In addition, we also compare prediction methods based on machine learning, such as random forests and neural networks, which have high prediction accuracy.

C0268: **Subclass discovery from fuzzy decision trees**

Presenter: **Christophe Marsala**, Sorbonne Universite, France

An approach based on decision trees and fuzzy sets theory is presented to highlight sub-classes in a set of classes. This approach is based on the use of fuzzy sets measures in order to determine non-connected subsets of classes. In a supervised learning approach, the aim is to find convenient features to determine an estimation of the decision frontier that separate different classes. In the case of a (fuzzy) decision tree, this estimation is built by means of a sequence of splits perpendicular to the feature axes. However, it often appears that leaves of the tree labeled with a similar class can be associated with either close regions of examples introducing connected regions of this class in the description space, or distinct and clearly separate regions of examples related to non-connected regions of examples labeled with similar class. This kind of approach is also known as subclass discovery. The aim is to propose an approach that combines a fuzzy decision tree and a clustering approach in order to highlight connected regions. Afterwards, this method is introduced in an explainable artificial intelligence (XAI) approach in order to propose a new way to build surrogates or to find counterfactual examples.

Authors Index